UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale

Corso di laurea magistrale in
Psicologia Cognitiva Applicata

Tesi di laurea magistrale

Il gambling attraverso la pubblicità: un'applicazione di statistica testuale

Text mining applied to gambling tv spots

Relatore
Prof. Calcagnì Antonio

Correlatore
Prof. Canale Natale

Laureando: Polo Niccolò

Matricola: 1154674

Anno Accademico 2018/2019

Indice

Introduzione

Con l'avvento di Internet e il sempre maggiore utilizzo della rete, enormi quantità di dati vengono accumulati in continuazione. Questa enorme disponibilità di materiale ha spronato lo sviluppo di una serie di competenze che permettessero il trattamento di grandi quantitativi di dati, i *Big Data*. Essi offrono un mondo sconfinato di possibilità di ricerca a cui attingono discipline come l'ingegneria infromatica, l'antropologia, la sociologia, la linguistica e molte altre ancora. Anche la psicologia può acquisire moltissime nuove informazioni dall'adeguato sfruttamento dei *Big Data*. Lo studio che verrà presentato nasce proprio da quest'idea: proporre un metodo alquanto inusuale per la ricerca psicologica, il *Text Mining*, che possa essere utilizzato sia per questo studio, sia essere esportato in altri contesti. L'indagine proposta si focalizza sullo studio del linguaggio adottato dalle pubblicità inerenti al gioco d'azzardo. Sono stati visualizzati tutti gli spot televisivi disponibili nel web andati in onda dal 2012 al 2018, e poi sono stati trascritti. Il dato testuale è un'informazione *sui generis* che deve essere elaborata con tecniche computazionali sviluppate *ad hoc*. Il *Natural Language Processing* (NLP) assolve a questa funzione, permettendo ai calcolatori di processare e trattare adeguatamente il linguaggio naturale. L'analisi è stata effettuata tramite l'adozione del processo di *Text Mining*, una tecnica di *Data Mining* che mediante l'utilizzo dell'NLP consente di estrapolare nozioni, individuare pattern e rappresentare visualizzazioni grafiche dei testi. È solo grazie all'analisi automatica dei testi che è possibile ricavare quelle informazioni dai testi che altrimenti rimarrebbero celate e non esplorabili.

Capitolo 1

1.1 Data mining

Il *Data Mining* è un processo che permette di esplorare grosse banche dati ed estrapolare informazioni e variabili tra esse. Tutto ciò è impensabile senza il prezioso supporto del computer e di software specifici. Il *Data Mining* insomma fa sì che da un'enorme mole di informazioni all'apparenza criptiche vengano estratti pattern, associazioni, schemi ricorrenti, in modo tale da poter giungere ad una conoscenza sfruttabile. L'intero processo prende il nome di *KDD*, un acronimo che sta a significare *Knowledge Discovery in Databases* (Camilo et al. 2015). La procedura del *KDD* consiste in una serie di passaggi sequenziali:

- Si identifica l'obiettivo da raggiungere: quali sono le ipotesi di ricerca che si intende indagare? Come si intende impostare il lavoro di ricerca?
- Vengono selezionati i dati attinenti all'obiettivo: utilizzare dati sufficienti e coerenti con il progetto di ricerca è fondamentale. Essi costituiscono le fondamenta su cui poggia l'impalcatura di tutto il progetto; materiale insufficiente o non adeguato può portare a non estrarre alcuna informazione rilevante o ancor peggio trarre conclusioni errate.
- I dati selezionati vengono puliti e poi sottoposti ad un pre-processamento e ad una pre-elaborazione: le informazioni vengono ripulite in modo tale da rimuovere il "rumore" che può essere causa di confusione nelle successive fasi di analisi. Poi i dati sono sottoposti ad una esplorazione preliminare e preparati per gli step successivi.
- I dati sono trasformati nel formato idoneo al software d'analisi utilizzato. La varietà dei dati viene ridotta, cercando al contempo di mantenere la qualità degli stessi.
- Segue la fase del data mining vero e proprio: si utilizza un software dedicato per scandagliare i dati raccolti e pre-pocessati. Vengono estratte le informazioni rilevanti attinenti all'obiettivo stabilito nel primo step.
- I risultati ottenuti vengono interpretati. Nel caso l'obiettivo non venga raggiunto è necessario ripetere lo step precedente, ed eventualmente anche quelli antecedenti.
- Visualizzazione dei risultati: i risultati vengono presentati in un formato che faciliti la comprensione e la divulgazione. I grafici sono un'eccellente metodo espositivo, chiaro e intuitivo.

1.2 Text Mining

Il *Text Mining* è una tecnica particolare di *Data Mining*, nata appositamente per trattare i dati di tipo testuale (Tan 1999). Il linguaggio naturale è per sua natura non strutturato e confuso: una caratteristica che

causa non pochi problemi all'applicazione della statistica sulle parole. Si potrebbe obiettare che la lingua scritta è tutt'altro che disordinata e scomposta, ma anzi segue regole linguistiche ben definite; inoltre il suo contenuto viene colto in maniera automatica con la semplice lettura. Ma questa abilità appartiene unicamente all'uomo: i calcolatori non sono in grado di comprendere il linguaggio scritto, i messaggi contenuti in esso, l'ironia, il sarcasmo, i modi di dire. Al contempo però applicare la statistica ai dati testuali permette di far emergere pattern, caratteristiche e informazioni nascoste che la lettura umana non è in grado di cogliere. Per applicare tecniche di *Text Mining* è necessario procedere in più passaggi, al pari della *KDD*. Di seguito verranno illustrate le fasi principali.

1.2.1 Acquisizione dati

Innanzitutto i dati vanno selezionati e raccolti in base ai propri obiettivi. Se lo scopo è estrarre informazioni di tipo economico è possibile scandagliare siti dedicati alla finanza o quotidiani che trattano l'economia. Se invece si vuole condurre indagini di tipo comportamentale su un largo campione di popolazione, i social network offrono una preziosa fonte di conoscenza a cui attingere. Questo processo prende i nome di *information retrieval* (Belkin and Croft 1992). L'*information retrieval* consiste in un insieme di paradigmi applicativi il cui scopo è quello di individuare le informazioni desiderate, memorizzarle e organizzarle in maniera coerente. In tal modo esse possono essere utilizzate e sfruttate in base alle necessità di chi le cerca. Due sono gli aspetti cruciali dell'*information retrieval*: la *query*, ovvero la stringa di parole chiavi utilizzata per scremare e selezionare le informazioni coerenti con l'indagine dell'utente. L'*oggetto*, ovvero i dati che dovrebbero corrispondere alla richiesta dell'utente. L'esempio più classico dell'applicazione dell'*information retrieval* è rappresentato dai motori di ricerca che tutti noi utilizziamo quotidianamente, come Google o Yahoo. Un utente interessato in un viaggio in Thailandia inserisce le parole chiavi "Thailandia", "viaggio", "soggiorno" (esse corrispondo alla *query*), il motore di ricerca scandaglia i propri server alla ricerca di informazioni coerenti con la richieste e restituisce una lista di siti web organizzati in maniera coerente (ovvero l'oggetto). Se il sistema di *information retrieval* è sviluppato in maniera efficiente esso risponderà in maniera congruente ai bisogni dell'utente. Al contrario, se le informazioni presentate non combaciano con le richieste, l'utente sarà insoddisfatto, producendo un danno per il gestore del motore di ricerca.

1.2.2 Pre-processamento

Il pre-processamento è probabilmente il passaggio più importante nell'analisi automatica dei testi, nonché il più complesso. Senza di esso il testo risulta essere un ammasso di informazioni sconclusionato e ridondante per i software di analisi testuale. Una volta individuata la documentazione che si desidera utilizzare per svolgere le proprie indagini si crea un *corpus*. Un *corpus* (*corpora* al plurale) è nient'altro che una raccolta di testi accuratamente organizzati. Esso è il materiale di base su cui vengono svolte tutte le necessarie operazioni per poter giungere all'estrazione di informazioni nascoste racchiuse al suo interno. Ad esempio, l'intera

raccolta di romanzi di Jane Austen può costituire un *corpus* su cui operare (il linguaggio di programmazione `R` (R Development Core Team 2008) mette a disposizione l'intera collezione di scritti dell'autrice tramite il pacchetto `janeaustenr` (Silge 2017)). Un sito di vendite online che voglia monitorare i giudizi dei clienti su uno specifico prodotto, potrebbe utilizzare come *corpus* la raccolta di tutte le recensioni del prodotto stesso. A seguire si effettua la *tokenizzazione*, il processo di divisione di caratteri in unità minime di analisi dette *token* (S and R 2016). I *token* sono sequenze di caratteri delimitati da spazi, segni di punteggiatura, simboli. Generalmente la *tokenizzazzione* viene effettuata per distinguere una parola dall'altra. Nulla vieta però di *tokenizzare* intere frasi o periodi, a seconda del tipo di analisi che si vuole condurre. Successivamente vengono eliminate le *stopwords*, parole inutili al fine dell'analisi testuale. Questo gruppo è costituito da una serie di termini onnipresenti in qualsiasi tipo di materiale testuale, come articoli, pronomi, preposizioni, congiunzioni, verbi ausiliari e via dicendo. La loro presenza rende i testi ampi e pesanti, senza contribuire in modo significativo alla raccolta di informazioni. Per facilitare la loro rimozione ci si appoggia a database contenenti le parole "inutili" di una lingua. Tramite una linea di codice si individuano le stesse parole presenti simultaneamente nelle raccolte e nei documenti, poi vengono rimosse. Chiaramente questi database non sono esaustivi per qualsiasi testo si voglia processare. Nel caso in cui una situazione del genere accada è necessario creare manualmente una lista *ad hoc* di *stopwords*. La punteggiatura non è di alcuna utilità al fine dell'analisi, ragion per cui va cancellata *in toto*. Al pari della punteggiatura, anche tutte le cifre presenti nei documenti non apportano alcun contributo all'esplorazione delle informazioni testuali. Le lettere in stampatello maiuscolo devono essere trasformate in stampatello minuscolo. Questa accortezza garantisce che parole identiche, ma originariamente scritte in formati diversi, vengano considerate come equivalenti: "Andrò" e "andrò" per i comuni lettori sono semanticamente identiche, ma per gli algoritmi di analisi non sono processi "intelligenti", consapevoli che queste parole hanno lo stesso significato. Con la lingua inglese è utilizzato di frequente lo *stemming* (Vijayarani, Ilamathi, and Nithya 2018), un processo mediante il quale vengono identificate le radici delle parole ed eliminati i suffissi. Ad esempio applicando lo stemming a parole dalla stessa desinenza come '*swim*', '*swims*', '*swimmed*', si otterrà solo il suffisso '*swim*': un notevole risparmio! L'italiano è una lingua altamente polisemica, per cui è caldamente sconsigliato applicare questo processo.

1.2.3 Natural Language Processing

Il *Natural Language Processing* (elaborazione del linguaggio naturale, in italiano) è un processo utilizzato dai computer mediante il quale le informazioni scritte o parlate del linguaggio naturale sono elaborate in automatico (King and Reinold 2014). É doveroso specificare la differenza tra *NLP* e *text mining*, poiché spesso si crea una certa confusione comprendere cos'è l'uno e cos'è l'altro: per *text mining* si intende una serie di passaggi consecutivi volti a trattare i dati testuali ed estrapolare informazioni; l'elaborazione del linguaggio naturale invece è un processo utilizzato nell'analisi automatica dei testi che svolge il preciso compito di supportare la macchina nella "lettura" e "comprensione" del testo. In pratica coloro che praticano *text mining*

utilizzano l'*NLP* per la lettura automatica del testo.

1.2.4 Analisi

Una volta che i dati sono stati ripuliti e accuratamente organizzati è possibile procedere con l'analisi vera e propria. In quest'ultima fase si visualizzano le statistiche descrittive di base relative ai documenti trattati, come i termini di maggiore utilizzo o il livello di ricchezza lessicale. É possibile creare dei *cluster*, ovvero applicare una tecnica di raggruppamento dei dati in base alle similarità o omogeneità degli elementi analizzati. In tal modo si identificano i principali topic presenti nei documenti. La modellazione è uno strumento molto efficace per effettuare previsioni su materiale testuale differente da quello preso in analisi. Attualmente è molto in voga la *sentiment analysis*, soprattutto per condurre indagini sulla percezione della politica o sull'andamento economico (Liu 2012) . Essa permette di assegnare dei valori indicativi della carica emotiva alle parole nei testi connessi ad un dato argomento. Così facendo si estraggono informazioni e si conducono indagini sulla sfera soggettiva.

Capitolo 2

2.1 Gioco d'azzardo, definizione e cenni storici

L'enciclopedia Treccani definisce il gioco d'azzardo "*un'attività ludica in cui ricorre il fine di lucro e nella quale la vincita o la perdita è in prevalenza aleatoria, avendovi l'abilità un'importanza trascurabile. Ne esistono svariati tipi, dai più antichi, come il gioco dei dadi (azzardo deriva dall'arabo az-zahr, che significa dado), a quelli più recenti effettuati con apparecchi automatici o elettronici. Possono dar luogo a una condizione patologica di dipendenza consistente nell'incapacità cronica di resistere all'impulso al gioco, con conseguenze anche gravemente negative sull'individuo stesso, la sua famiglia e le sue attività professionali*". Sono tre le caratteristiche del gioco su cui l'enciclopedia pone l'accento: il possibile guadagno in denaro, il motore che spinge le persone a intraprendere questa attività; l'aleatorietà delle vincite, a cui troppo spesso non viene dato il giusto peso in favore dell'erronea convinzione che l'abilità e l'esperienza personale possano far pendere la bilancia delle vincite in proprio favore; il rischio patologico di sviluppare una dipendenza altamente dannosa per sé e per i familiari. Il gioco d'azzardo è un'attività umana praticata fin dagli albori della civiltà. Scavi archeologici hanno rinvenuto tracce di questa pratica già millenni or sono. I primi dadi sono stati recuperati in Cina e risalgono a 5000 anni fa; altri reperti sono stati raccolti in Egitto, India, Giappone. Nell'Antica Grecia, il poeta Sofocle ha affermato che i dadi sono stati inventati da un eroe mitologico durante l'assedio di Troia e, sebbene questo possa avere una base piuttosto dubbia, i suoi scritti intorno al 500 a.C. sono stati la prima menzione dei dadi della storia greca. É noto che il gioco d'azzardo fosse largamente praticato anche all'epoca dell'Impero Romano. Per ragioni di ordine pubblico era vietato dedicarsi a questo passatempo, divieto che comunque non ha impedito che venisse praticato. Era però consentito scommettere, puntando sulle corse delle bighe e delle quadrighe. Sempre all'epoca dei romani fu inventato il precursore di quello che oggi chiamiamo roulette. All'interno delle legioni infatti era usanza far roteare uno scudo sopra una lancia come una rudimentale roulette. Solo nel XVII secolo fu perfezionata da Blaise Pascal, che ha dato forma al gioco come lo conosciamo noi oggi. Sempre alla Cina spetta il primato dell'invenzione delle carte gioco, datato intorno al IX secolo. Una cosa però è certa: le carte usate in questa epoca storica hanno una scarsa relazione con le carte standard da 52 pezzi conosciute oggi. Da questo momento in poi il gioco d'azzardo ha subito una svolta sicuramente importante e tra i giochi che più di tutti hanno contribuito a al cambiamento c'è sicuramente il Blackjack. Alcuni suggeriscono che le prime forme di Blackjack venivano da un gioco spagnolo chiamato *ventiuna* (letteralmente "ventuno"), visto che questo gioco apparve in un libro scritto dall'autore di Don Chisciotte nel 1601. Inoltre il gioco francese di *vingt-et-un*, nel XVII secolo, è certamente un predecessore diretto del gioco moderno. Anche se per arrivare al nome che ha ai giorni nostri bisogna attendere fino al secolo scorso, poiché il nome di 'blackjack' fu un'innovazione americana legata a delle promozioni speciali nei casinò del Nevada negli anni '30. Per il gioco "principe" di tutti i casinò, il poker, risulta però difficile individuarne l'origine precisa. Alcuni hanno notizie di poker provenienti dalla Persia del XVII secolo, mentre

altri dicono che il gioco che conosciamo oggi è stato ispirato da un gioco francese chiamato Poque. Quello che sappiamo con certezza è che un attore inglese con il nome di Joseph Crowell ha riferito che una forma riconoscibile del gioco è stata giocata a New Orleans nel 1829, perciò è una buona data per la nascita del poker. Nel XV secolo in Italia nacquero le prime lotterie. Più precisamente il gioco del Lotto viene fatto risalire a Venezia nel 1734. I guadagni venivano incassati dalla città stessa, e i cittadini che giocavano non si rendevano conto che sostanzialmente stavano giocando a pagare una nuova tassa. L'età d'oro del gioco d'azzardo comincia negli anni '90, grazie alla progressiva diffusione della rete e di strumenti digitali come pc, tablet e smartphone. Il web ha permesso di troncare di netto le distanze tra giocatore e sale da gioco. Seduti comodamente a casa è possibile cimentarsi in tornei di poker con persone di tutto il mondo, tentare la fortuna con un giro di roulette o cercare di vincere contro il banco in una partita di blackjack, quasi sempre con esito negativo. Il gioco virtuale, a differenza di quello fisico, utilizza strumenti di seduzione alquanto insidiosi. Quasi tutte le piattaforme virtuali consentono al neofita di prendere parte al gioco gratuitamente. In aggiunta sono garantiti bonus in denaro all'iscrizione: non solo le persone sono indotte a prendere parte ad un'attività che comporta una perdita di denaro pressochè certa, ma sono anche abbindolati dall'idea di aver già vinto ancor prima di aver iniziato a giocare.

2.2 Regolamentazione

A livello comunitario non esiste alcuna regolamentazione specifica per il gioco d'azzardo. Una deliberazione degna di nota risale al 2013. Essa tratta la legittimità che ciascuno stato ha relativamente alla possibilità di intervenire in tutela dei giocatori, anche a discapito della libertà d'impresa dei distributori di giochi d'azzardo (*Corte di Giustizia, Sentenza del 22/01/2015, Causa Stanley International Betting Ltd e a. c. Ministero dell'Economia e delle Finanze, in relazione alla libera prestazione di servizi - giochi d'azzardo - sistema delle concessioni*). Nel 2014 sempre la Commissione Europea ha raccomandato agli stati membri di tutelare i consumatori, in particolare minori e soggetti a rischio, sui rischi del gioco online. Si sottolinea la necessità di fornire informazioni ai giocatori circa i rischi cui vanno incontro, di realizzare una pubblicità responsabile, di vietare ai minori l'accesso al gioco d'azzardo online, di creare un conto di gioco per determinare l'identità e, soprattutto, l'età del consumatore, con fissazione di un limite di spesa e messaggi periodici su vincite e perdite realizzate (*Commissione europea Comunicato stampa Bruxelles, 14 luglio 2014 Gioco d'azzardo on-line: la Commissione raccomanda principi intesi a tutelare efficacemente i consumatori*). Per quanto concerne l'Italia, la legislazione riguardo le concessione è fatta risalire al 1993, con diversi aggiornamenti nel corso degli anni (*legge finanziaria per il 2006, art.1, commi 525 ss, legge finanziaria per il 2006, art.1, commi 525 ss*). Essa prevede che la regolamentazione del settore sia affidata all'Azienda Autonoma Monopoli di Stato e relative sanzioni a chiunque non si attenga alle norme. La legge n. 88 del 2009, art. 24, commi 12 ss (*legge comunitaria per il 2008*), oltre a nuovi requisiti dei soggetti che richiedono la concessione ed un inasprimento delle sanzioni, prevede l'adozione di strumenti ed accorgimenti per l'esclusione dall'accesso

al gioco online da parte di minori, nonché l'esposizione del relativo divieto in modo visibile negli ambienti virtuali di gioco gestiti dal concessionario (*comma 17, lett. e*). Con la legge n. 220 del 2010 (*art. 1, commi 78 ss*) viene rivisto lo schema di convenzione tipo per le concessioni per l'esercizio e la raccolta dei giochi pubblici, anche al fine di contrastare la diffusione del gioco irregolare o illegale in Italia e le infiltrazioni della criminalità organizzata nel settore, di tutelare la sicurezza, l'ordine pubblico ed i consumatori, specie minori d'età (*sulla legittimità di tali restrizioni all'attività di organizzazione e gestione dei giochi pubblici affidati in concessione vedi anche la sentenza della Corte costituzionale n. 56 del 2015+*). Il decreto Balduzzi (*decreto legge n. 158 del 2012, convertito nella legge n. 189 del 2012*) affronta più tematiche relative al gioco in maniera più sostanziale. In particolare prevede l'aggiornamento dei livelli essenziali di assistenza (LEA) con riferimento alle prestazioni di prevenzione, cura e riabilitazione rivolte alle persone affette da ludopatia (art. 5, comma 2). Il decreto ribadisce l'importanza di dichiarare in modo esplicito il rischio di dipendenza cui il giocatore va incontro, e il divieto tassativo di far partecipare i minori al gioco. Per contenere i messaggi pubblicitari, si vieta l'inserimento di spot promozionali dei giochi con vincite in denaro nelle trasmissioni televisive e radiofoniche, durante le rappresentazioni teatrali e cinematografiche non vietate ai minori. Sono anche proibiti i messaggi pubblicitari di giochi con vincite in denaro che incitano al gioco, che ne esaltano la sua pratica, che hanno al loro interno dei minori, o che non avvertono del rischio di dipendenza dalla pratica del gioco. La pubblicità deve riportare in modo chiaramente visibile la percentuale di probabilità di vincita che il soggetto ha nel singolo gioco. In base al decreto Balduzzi è stato istituito infine un Osservatorio per valutare le misure più efficaci per contrastare la diffusione del gioco d'azzardo e il fenomeno della dipendenza grave. Questo decreto è un elemento chiave da tenere in conto al fine dell'analisi e comprensione dei dati presi in considerazione dalla seguente ricerca. Il decreto infatti pone dei paletti alquanto restrittivi su cosa e come possono esprimersi le pubblicità sul gioco d'azzardo. Altri decreti sono stati presentati nel corso degli anni successivi (*legge n. 190 del 2014, legge n. 208 del 2015*). Segnalo solo l'ultimo provvedimento (*decreto-legge n. 87 del 2018, convertito nella legge n. 96 del 2018*) che introduce il divieto assoluto di pubblicità dei giochi d'azzardo ed altre disposizioni per il contrasto dei disturbi da gioco d'azzardo.

2.3 Gioco d'azzardo e patologia

Il gioco d'azzardo è stato per lungo tempo oggetto di studi secondo diversi approcci psicologici, e tutt'ora è ancora oggetto di speculazioni. Freud, e con lui la scuola psicanalitica, ritengono che la causa del gioco d'azzardo compulsivo sia dovuto ad un sentimento di colpevolezza che debba essere espiato; il desiderio inconscio di perdere permette di ottenere un sollievo da questo stato di malessere (Freud 1928). Anche Bergler è in linea con questa ipotesi (Bergler 1957); per l'autore il giocatore cronico che persiste nelle scommesse nonostante le continue perdite è un nevrotico. Seppur egli dichiari di voler vincere e di aspettarsi di vincere, ciò che inconsciamente lo motiva è il desiderio di perdere e di essere punito. Il tutto è riconducibile ad una mancata risoluzione dei sentimenti infantili di onnipotenza e controllo onnipotente sul mondo e sul fato. È

di ben altra opinione la scuola comportamentale. Skinner ipotizza che il gioco d'azzardo compulsivo sia il risultato della formazione di un comportamento maladattivo (Skinner 1953). La dipendenza è causata dal rapporto tra vincite e perdite studiato appositamente per infondere un rinforzo positivo costante nel giocatore, che a lungo andare si cronicizza. Gli studi fisiologici si sono concentrati sullo studio del comportamento cerebrale di soggetti sottoposti a sessioni di gioco, tramite tecniche di imaging, in particolare l'fMRI (*functional magnetic resonance imaging*). È emerso che nei soggetti predisposti il gioco d'azzardo causa un'elevata attivazione del circuito della ricompensa (Jabr 2013). Tale circuito, se attivato, rilascia piccoli quantitativi di dopamina che provocano una sensazione di piacere e benessere. Il rilascio costante di dopamina a chi gioca d'azzardo fa insorgere una vera e propria dipendenza, al pari di chi fa utilizzo di droghe come cocaina o anfetamine. Al contempo il gioco è associato ad un aumento dell'arousal che si manifesta con un aumento della frequenza cardiaca e un incremento dei livelli di cortisone (Anderson and Brown 1984) (May et al. 2003). Teorie di psicologia della personalità invece evidenziano la presenza di un tratto tratto di propensione al rischio che permette distinguere i giocatori patologici rispetto ai giocatori non patologici (M. D. Griffiths 1990). Il gioco patologico nelle sue forme più estreme è una diagnosi psichiatrica riconosciuta nel Manuale di statistica e Diagnostica , versione 5 (American Psychiatric Association 2013). La severità di questa patologia è da intendere come continuum, e non come una specifica condizione generalizzabile a tutti coloro di cui ne sono affetti. Il paradosso del gioco risiede nel fatto che la netta maggioranza di chi lo pratica è ben conscio del fatto che i distributori guadagnano di più rispetto ai fruitori: detto volgarmente, tutti sanno che "il casinò vince sempre" (Clark 2010). Eppure, pur di fronte ad un saldo negativo certo, i giocatori non accennano a diminuire. Questo fenomeno è spiegabile dal fatto che la motivazione al gioco non è dettata da una valutazione economica logica e razionale, ma esistono fattori emotivi e cognitivi che inducono a perseverare in questa pratica. Gli studi di psicologia cognitiva condotti sul gioco d'azzardo hanno fatto emergere sistematiche distorsioni cognitive presenti tanto nei giocatori occasionali quanto in quelli abituali. Le persone tendono regolarmente a sovrastimare le loro probabilità di vincita (Sharpe 2002). Un enorme contributo alla psicologia del gioco d'azzardo viene dagli esperimenti di Langer riguardo l'illusione del controllo (Langer 1975). Questa distorsione consiste nella sovrastima delle proprie possibilità di successo rispetto alle reali probabilità di vittoria. In giochi che danno ampio spazio alle decisioni sulle azioni di gioco, come il poker o il blackjack, le persone tendono erroneamente a credere di essere capaci o più bravi rispetto agli avversari; questa eccessivo senso di sicurezza induce a scommettere più spesso e con maggiori somme di denaro, e di conseguenza le perdite risultano essere maggiori. Paradossalmente l'*Illusion of Control* è presente anche per le lotterie e i gratta e vinci, per i quali le decisioni che si compiono (comprare un biglietto rispetto ad un altro) non hanno il ben che minimo impatto sull'esito del gioco. Lang ha riscontrato in una serie di esperimenti che chi acquista un biglietto della lotteria è reticente a scambiarlo con un altro biglietto. Addirittura l'acquirente attribuisce al proprio biglietto un valore in denaro superiore rispetto ad un altro del tutto identico, proprio perché la scelta del biglietto lo investe di un'importanza superiore rispetto a quelli non scelti. Reid scopre che i fallimenti molto vicini ad un successo motivano ed incoraggiano le persone a continuare a giocare (Reid

1986). Il fenomeno prende il nome di *Near Miss Effect*. Reid ritiene che il quasi successo ha lo stesso effetto condizionante di un vero e proprio successo, e questo rinforza la motivazione al gioco. Le slot machine approfittano proprio di questa distorsione cognitiva mostrando con alta frequenza combinazioni quasi complete di simboli identici, in modo da fornire l'impressione di aver "quasi vinto". Secondo Kahneman e Tversky il *Near Miss Effect* è spiegato dalla frustrazione della vittoria mancata per poco, o come la definiscono loro, *Cognitive Regret* (Kahneman and Tversky 1982) . La frustrazione viene neutralizzata decidendo di giocare di nuovo (G. R. Loftus and Loftus 1983).

Capitolo 3

3.1 Ipotesi di ricerca

La ricerca prende spunto da uno studio condotto da un gruppo di ricerca allo scopo di analizzare l'approccio comunicativo che adottano le aziende di scommesse sportive per indurre i potenziali consumatori a fruire del gioco d'azzardo (Lopez-Gonzalez, Guerrero-Solé, and Griffiths 2018). Gli studiosi hanno raccolto una serie di pubblicità riguardanti le scommesse sportive calcistiche trasmesse tra il 2014 e il 2016 nel Regno Unito e in Spagna, poi sono state identificate una serie di categorie in cui potessero essere inclusi dei pattern di comportamento presentati nei video pubblicitari. L'approccio di *ricerca* utilizzato è quello della *content analysis*. Lo scopo della ricerca era quello di individuare quali fossero le situazioni suggerite dalle pubblicità che portino la popolazione a ritenere il gioco d'azzardo come un comportamento socialmente accettabile e del tutto normale da praticare. Il quadro teorico su cui poggia lo studio è la *Teoria delle Rappresentazioni Sociali* (S Moscovici 2015), secondo la quale le rappresentazioni sociali condivise hanno un duplice scopo: innanzitutto *convenzionalizzano* oggetti, individui ed eventi dando loro una forma definita attraverso la loro ripetizione in molteplici contesti di rappresentazioni sociali; in secondo luogo invitano coloro cui sono soggetti a quei specifici contesti a tenere una serie di comportamenti in linea con quanto sono rappresentati i suddetti contesti (Serge Moscovici 1984). l'ipotesi di ricerca che verrà illustrata di seguito attinge agli studi di *Tizio e Caio* ma sposta il suo focus di indagine: le informazioni analizzate non sono di tipo visivo, bensì di tipo testuale. Sfruttando le tecniche di *Text Mining* e *Natural Language processing* l'intenzione è quella di esplorare le informazioni contenute all'interno dei messaggi verbali delle pubblicità televisive riguardanti tutti i giochi d'azzardo (quindi non solo le scommesse sportive calcistiche) trasmesse nel piccolo schermo dal 2012 al 2018.

3.2 Raccolta dati

I dati selezionati sono costituti dai testi delle pubblicità riguardanti il gioco d'azzardo trasmesse in televisione tra il 2012 e il 2018. La selezione dei *brand* è avvenuta tramite la consultazione di blog e siti specializzati. Tutti i *brand* scelti hanno la licenza di operare legalmente nel mercato italiano concessa dall'Agenzia delle Dogane e dei Monopoli. Le pubblicità sono state visionate sulla piattaforma di condivisione video *Youtube*. I testi sono stati trascritti manualmente tramite il supporto, ove possibile, del sito *DIYCaptions*. Il numero totale di pubblicità trascritte ammonta a 160. Esse sono state suddivise in tre *corpora*: 45 spot sono stati raggruppati sotto la categoria di pubblicità di scommesse sportive; 55 spot sono stati raggruppati nella categoria di giochi in stile lotto, tombola, gratta e vinci; i restanti 60 spot sono stati classificati come appartenenti alla categoria di giochi in stile casinò. La scelta di formare queste tre categorie in parte è basata sulla classificazione delle diverse tipologie di gioco da parte dell'Agenzia delle Dogane e dei Monopoli, in parte è dettata dalle

caratteristiche in comune che hanno i diversi giochi. Le categorie 'scommesse sportive' e 'giochi in stile casinò' combaciano con le catalogazioni dell'Agenzia sotto il nome rispettivamente di 'giochi a base sportiva' e 'Giochi di abilità, Carte, Sorte a quota fissa'. Invece si è deciso di raggruppare sotto la terza categoria i giochi catalogati dall'Agenzia come 'Giochi numerici a quota fissa', 'Giochi numerici a totalizzatore' e 'Bingo', in quanto questi giochi presentano modalità di interazione estremamente simili tra loro.

3.3 Metodo

Di seguito verrà presentata la procedura adottata per sfruttare l'analisi testuale a scopo esplorativo. Saranno presentati passo passo i processi utilizzati con annessi sia gli output che le linee di codice utilizzate, in modo tale da favorire la comprensione e la replicabilità dello studio, il tutto in un ottica di *Open Science*. Il lavoro è stato svolto dapprima su notebook *ASUS X54C* e poi su notebook *Toshiba T410* a causa di un imprevisto tecnico (leggasi CPU eccessivamente surriscaldata). Il sistema operativo utilizzato è *Linux* con distribuzione *Xubuntu*. Il linguaggio di programmazione che ha consentito le analisi è `R`, supportato dall'*IDE* (*Integrated Development Enviroment* - Ambiente di Sviluppo Integrato) *RStudio*, entrambi gratuiti e *open-source*. La versione di R adottata è stata inizialmente la 3.5.2 (Eggshell Igloo), poi aggiornata alla versione 3.6.0 (Planting of a Tree).

3.3.1 Pacchetti

Per poter operare una serie di processi avanzati su R è necessario installare e caricare progressivamente dei pacchetti preposti a svolgere specifiche funzioni utili al lavoro di Text Mining e Natural Language Processing. Il lavoro è stato svolto principalmente grazie all'ausilio del pacchetto `quanteda`(Benoit et al. 2018), che comprende una vasta gamma di funzioni specifiche per il trattamento del linguaggio naturale e l'analisi testuale. Un altro pacchetto dalle funzioni simili è `tm`(Feinerer, Hornik, and Meyer 2008), che è stato utilizzato parzialmente per svolgere alcune operazioni non coperte da `quanteda`. Altri pacchetti sono stati utilizzati per svolgere funzioni non direttamente attinenti all'elaborazione testuale, ad esempio rappresentazioni grafiche, concatenazione delle linee di codice, tagging. In ordine di utilizzo essi sono:`readtext` (Benoit and Obeng 2019), `tidytext` (Silge and Robinson 2016), `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2019), `knitr` (Xie 2014), `lubridate` (Grolemund and Wickham 2011), `topicmodels` (Grün and Hornik 2011), `heatmap3` (S. Zhao et al. 2019), `cluster` (Maechler et al. 2019), `dendextend` (Galili 2015), `udpipe` (Wijffels 2019), `lattice` (Sarkar 2008), `gridExtra` (Auguie 2017), `syuzhet` (Jockers 2015).

3.3.2 Caricamento dati

I dati (ovvero i testi) sono stati salvati in formato .txt; i file prendono il nome `poker.txt` (giochi in stile casinò), `sport.txt` (scommesse sportive) e `lotto.txt` (tombola, lotto, gratta e vinci). Essi, per poter essere

elaborati, devono essere prima importati tramite la funzione `readtext` che consente il trattamento di dati di tipo testuale.

```
setwd("/home/niccolo/Scrivania/TESI/tesi/txt/")
poker <- readtext("/home/niccolo/Scrivania/TESI/tesi/txt/poker.txt")
sport <- readtext("/home/niccolo/Scrivania/TESI/tesi/txt/sport.txt")
lotto <- readtext("/home/niccolo/Scrivania/TESI/tesi/txt/lotto.txt")
```

3.3.3 Preprocessamento

Attraverso il preprocessamento i dati sono ripuliti dal rumore, ovvero le informazioni superflue e ridondanti. Innanzitutto i testi caricati sono trasformati in *corpus*, l'oggetto base su cui operano le funzioni del pacchetto `quanteda`.

```
poker1 <- corpus(poker)
sport1 <- corpus(sport)
lotto1 <- corpus(lotto)
```

Inoltre viene creato un corpus comprensivo di tutti e tre i *corpora*.

```
azzardo <- poker1 + sport1 + lotto1
summary(azzardo)
```

```
Corpus consisting of 3 documents:

      Text Types Tokens Sentences
 poker.txt   807   3286       342
 sport.txt   762   2298       106
 lotto.txt  1028   4039       383

Source: Combination of corpuses poker1 + sport1 and lotto1
Created: Thu Jun 20 09:20:49 2019
Notes:
```

Le *stopwords* utilizzate non appartengono al dizionario italiano incluso in `quanteda`. Si è scelto invece di creare un file di testo esterno (N=701) contentente tutte le *stopwords* tipiche dell'italiano con l'aggiunta di ulteriori parole come i nomi propri delle aziende fornitrici di gioco d'azzardo. In questo modo il vocabolario di parole da rimuovere risulta essere altamente specifico per i propri testi da ripulire.

```
stopwords1 <- readLines("/home/niccolo/Scrivania/TESI/tesi/txt/stopwords.txt",
                        encoding = "UTF-8")
```

La pulizia vera e propria è effettuata distintamente per ciascun *corpus*. Il processo prevede una serie di step sequenziali. Il *corpus* è suddiviso in *tokens*, stringhe di caratteri corrispondenti ad ogni parola delimitate da alcuni demarcatori (linee di spazio, punteggiatura, simboli). Poi si esegue una rimozione di numeri, punteggiatura e simboli. I caratteri in stampatello maiuscolo sono trasformati in stampatello minuscolo. Infine sono rimosse le parole coincidenti a quelle contenute nel dizionario di stopwords caricato nello step precedente.

```
poker3 <- poker1 %>% tokens(remove_numbers = TRUE,
                            remove_punct = TRUE,
                            remove_symbols = TRUE) %>%
                     tokens_tolower() %>%
                     tokens_select(stopwords1,
                                   selection = "remove",
                                   padding = FALSE,
                                   verbose = TRUE) %>%
                     tokens_select("può",
                                   selection = "remove",
                                   padding = FALSE,
                                   verbose = TRUE)
```

```
removed 261 features
removed 1 feature
```

```
sport3 <- sport1 %>% tokens(remove_numbers = TRUE,
                            remove_punct = TRUE,
                            remove_symbols = TRUE) %>%
                     tokens_tolower() %>%
                     tokens_select(stopwords1,
                                   selection = "remove",
                                   padding = FALSE,
                                   verbose = TRUE)
```

```
removed 237 features
```

```
lotto3 <- lotto1 %>% tokens(remove_numbers = TRUE,
                            remove_punct = TRUE,
```

```
                              remove_symbols = TRUE) %>%
                    tokens_tolower() %>%
                    tokens_select(stopwords1,
                                  selection = "remove",
                                  padding = FALSE,
                                  verbose = TRUE)
```

```
removed 272 features
```

```
azzardo3 <- azzardo %>% tokens(remove_numbers = TRUE,
                               remove_punct = TRUE,
                               remove_symbols = TRUE) %>%
                        tokens_tolower() %>%
                        tokens_select(stopwords1,
                                      selection = "remove",
                                      padding = FALSE,
                                      verbose = TRUE) %>%
                        tokens_select("può",
                                      selection = "remove",
                                      padding = FALSE,
                                      verbose = TRUE)
```

```
removed 408 features
removed 1 feature
```

```
head(poker3[[1]],20)
```

```
 [1] "casinò"       "fiato"        "sospeso"      "inizia"
 [5] "giocare"      "gratis"       "giocate"      "regalo"
 [9] "gioco"        "oggi"         "modo"         "entrare"
[13] "casinò"       "online"       "divertimento" "aspetta"
[17] "pc"           "smartphone"   "tablet"       "scarica"
```

Come si può notare dal *chunk* di codice, solo il file `poker1` ha subito la rimozione della forma verbale 'può'. La scelta di cancellare il verbo da quell'oggetto è stata presa in seguito ad un'analisi di contesto operata proprio sulla parola 'può'. Generalmente il verbo 'potere' e tutte le sue forme verbali sono catalogate come *stopwords*, e di conseguenza eliminate. Ma dal momento che si sta operando su testi pubblicitari, si è ritenuto che il valore di potenzialità dato da alcuni verbi ai messaggi degli spot fosse un tassello importante al fine dell'analisi. Si può presumere che il verbo 'potere' possa essere sfruttato per indurre il consumatore a credere

che le proprie abilità incidano grandemente sull'eventuale vincita in denaro (es. Gioca a -nome brand-, **puoi** vincere fino a 10000 euro). Il codice ora presentato mostra uno stralcio di analisi contestuale della parola *target*.

```
contesto_può <- kwic(azzardo,
                     pattern = 'può',
                     window = 4)
head(contesto_può, 10)
```

```
  [poker.txt, 28]   vietato ai minori e | può |
 [poker.txt, 104]          it. Il gioco | può |
 [poker.txt, 161]   vietato ai minori e | può |
 [poker.txt, 228]   vietato ai minori e | può |
 [poker.txt, 284] maggiorenni. Il gioco | può |
 [poker.txt, 347] maggiorenni. Il gioco | può |
 [poker.txt, 452] maggiorenni. Il gioco | può |
 [poker.txt, 512] maggiorenni. Il gioco | può |
 [poker.txt, 607] maggiorenni. Il gioco | può |
 [poker.txt, 672]   vietato ai minori e | può |

 causare dipendenza. Da
 causare dipendenza patologica.
 causare dipendenza patologica Sei
 causare dipendenza."
 causare dipendenza patologica.
 causare dipendenza patologica.
 causare dipendenza patologica.
 causare dipendenza patologica.
 causare dipendenza patologica.
 causare dipendenza patologica.
```

Si noti che per quanto riguarda il *corpus* `poker`, il verbo 'può' è utilizzato unicamente all'interno dei messaggi di avvertenze presenti secondo obbligo di legge. Dal momento che gli avvisi di pericolo non sono dei concetti espressi per volontà delle aziende ma per obbligo normativo, si è deciso che tali informazioni fossero irrilevanti e perciò rimosse. In conclusione si creano le matrici `DFM` (*document feature matrix*). Esse sono delle strutture ampiamente utilizzate per l'analisi nel *Text Mining* in cui ogni riga rappresenta un documento, ogni colonna rappresenta un lemma e ogni valore corrisponde alla frequenza di comparsa del lemma nel documento.

```
poker4 <- dfm(poker3, stem = FALSE)
sport4 <- dfm(sport3, stem = FALSE)
lotto4 <- dfm(lotto3, stem = FALSE)
azzardo4 <- dfm(azzardo3, stem = FALSE)
head(azzardo4, n = 3, nf = 8)
```

```
Document-feature matrix of: 3 documents, 8 features (41.7% sparse).
3 x 8 sparse Matrix of class "dfm"
           features
docs        casinò fiato sospeso inizia giocare gratis giocate regalo
  poker.txt     55     1       1      2      10      8       6      1
  sport.txt      0     0       0      0       9      1       1      0
  lotto.txt      0     0       0      2      25      0       0      1
```

3.3.4 Indici descrittivi

Una volta ripuliti i testi si procede con l'esplorazione delle informazioni contenute in essi. Un primo indice molto utilizzato è il *Type Token Ratio*. Come suggerisce il nome esso è il rapporto tra i *types* (il numero totale di parole differenti) e le *tokens* (il numero totale di parole), e restituisce un indice di diversità lessicale che varia da 0 a 1. Più il punteggio si avvicina a 1, maggiore è la ricchezza lessicale e viceversa.

```
sapply(c(poker4,sport4,lotto4), textstat_lexdiv)
```

```
         [,1]         [,2]         [,3]
document "poker.txt" "sport.txt" "lotto.txt"
TTR      0.4507317   0.5773196   0.4573477
```

Se si prende a riferimento la media di punteggio per il corpus di riferimento dell'italiano parlato in TV (0,47) (Spina and Umane 2014), appare evidente che la diversità lessicale dei tre corpora è piuttosto scarsa. Questo è indicativo del fatto che le pubblicità utilizzano una comunicazione verbale piuttosto concisa, diretta e semplice da comprensere. È possibile studiare la diversità lessicale anche estrapolando il numero di *hapax* (parole che ricorrono una sola volta) contenuti nei testi.

```
rowSums(azzardo4 == 1) %>% head()
```

```
poker.txt sport.txt lotto.txt
      327       352       415
```

```
((rowSums(azzardo4 == 1))/rowSums(azzardo4))*100 %>% head()
```

```
poker.txt sport.txt lotto.txt
```

```
 31.90244  40.45977  29.85612
```

Si noti come il documento **sport** presenta una percentuale di hapax superiore agli altri due testi e correla con il valore TTR più elevato.

3.3.4 Analisi di frequenza

Una volta estratti gli indici di diversità lessicale si è proceduto allo studio delle parole in base alla loro frequenza di comparsa nei testi. Procedendo in ordine per ciascun documento, per prima cosa vengono **estratte** i primi venti termini a maggiore frequenza.

```
textstat_frequency(azzardo4, n = 5, groups = c("poker","sport","lotto"))
```

```
     feature frequency rank docfreq group
1       euro        36    1       1 lotto
2       puoi        28    2       1 lotto
3    vincere        28    2       1 lotto
4     numeri        28    2       1 lotto
5    giocare        25    5       1 lotto
6      gioco        73    1       1 poker
7     casinò        55    2       1 poker
8       euro        31    3       1 poker
9      bonus        23    4       1 poker
10     poker        18    5       1 poker
11 giocatori        26    1       1 sport
12     bonus        16    2       1 sport
13     quote        16    2       1 sport
14      euro        13    4       1 sport
15     gioca        11    5       1 sport
```

Le rappresentazioni grafiche consentono di comprendere più agilmente la distribuzione dei termini più significativi all'interno di ciascun testo.

```
pok_freq<- poker4 %>% dfm_trim(min_termfreq = 3) %>%
  textstat_frequency()
pok_freq$feature<-with(pok_freq, reorder(feature,-frequency))
ggplot(pok_freq, aes(x = feature, y = frequency)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 8))
```

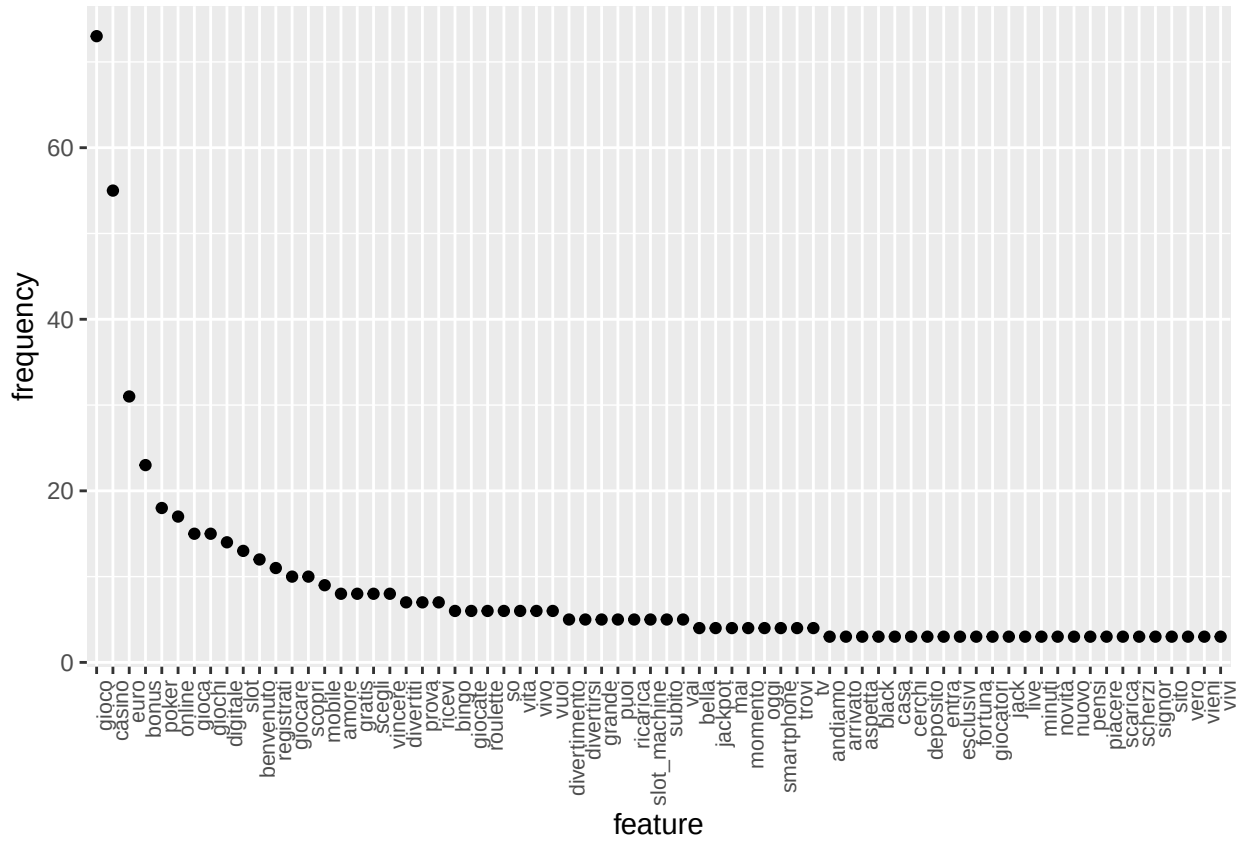


Figure 1: Distribuzione frequenze superiori a 2 testo poker

```
spo_freq<- sport4 %>% dfm_trim(min_termfreq = 3) %>%
  textstat_frequency()
spo_freq$feature<-with(spo_freq, reorder(feature,-frequency))
ggplot(spo_freq, aes(x = feature, y = frequency)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 8))
```

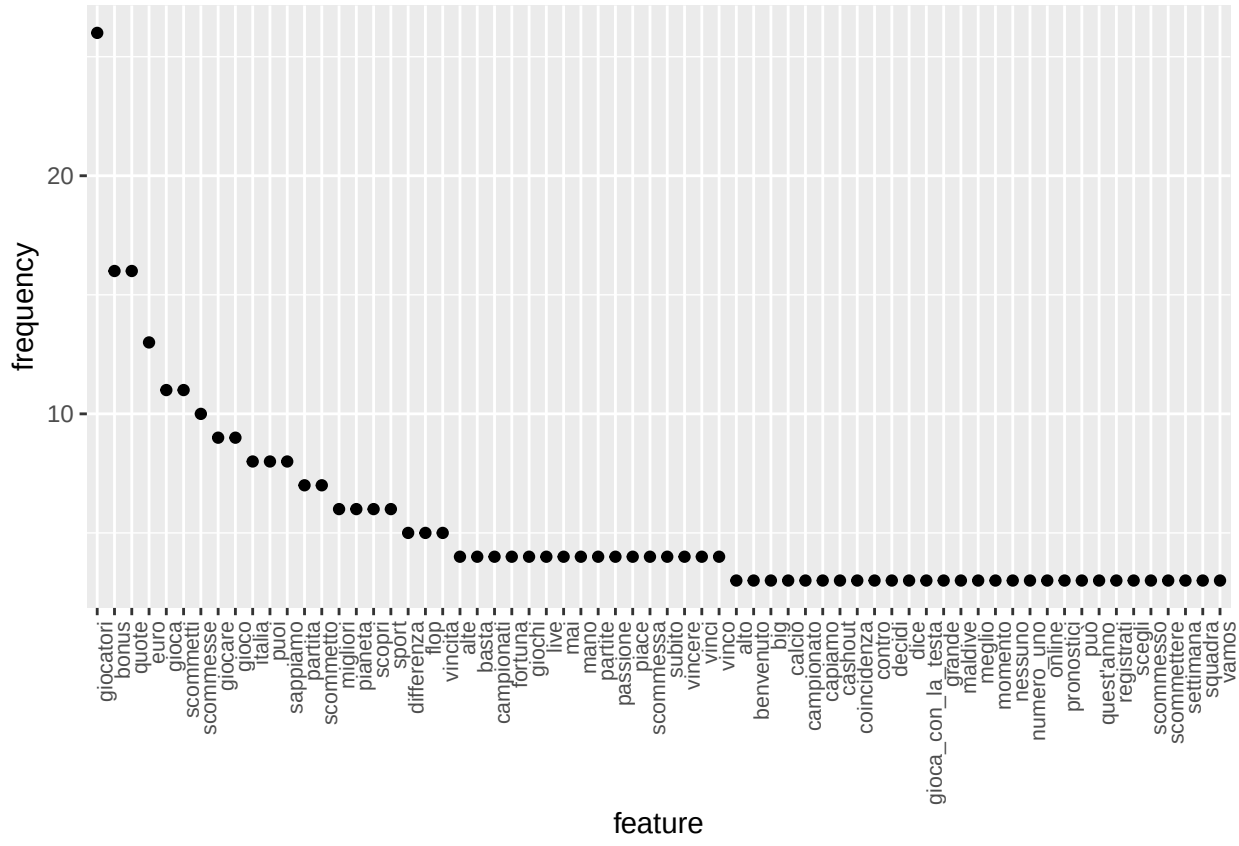


Figure 2: Distribuzione frequenze superiori a 2 testo sport

```
lot_freq<- lotto4 %>% dfm_trim(min_termfreq = 4) %>%
  textstat_frequency()
lot_freq$feature<-with(lot_freq, reorder(feature,-frequency))
ggplot(lot_freq, aes(x = feature, y = frequency)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 6))
```

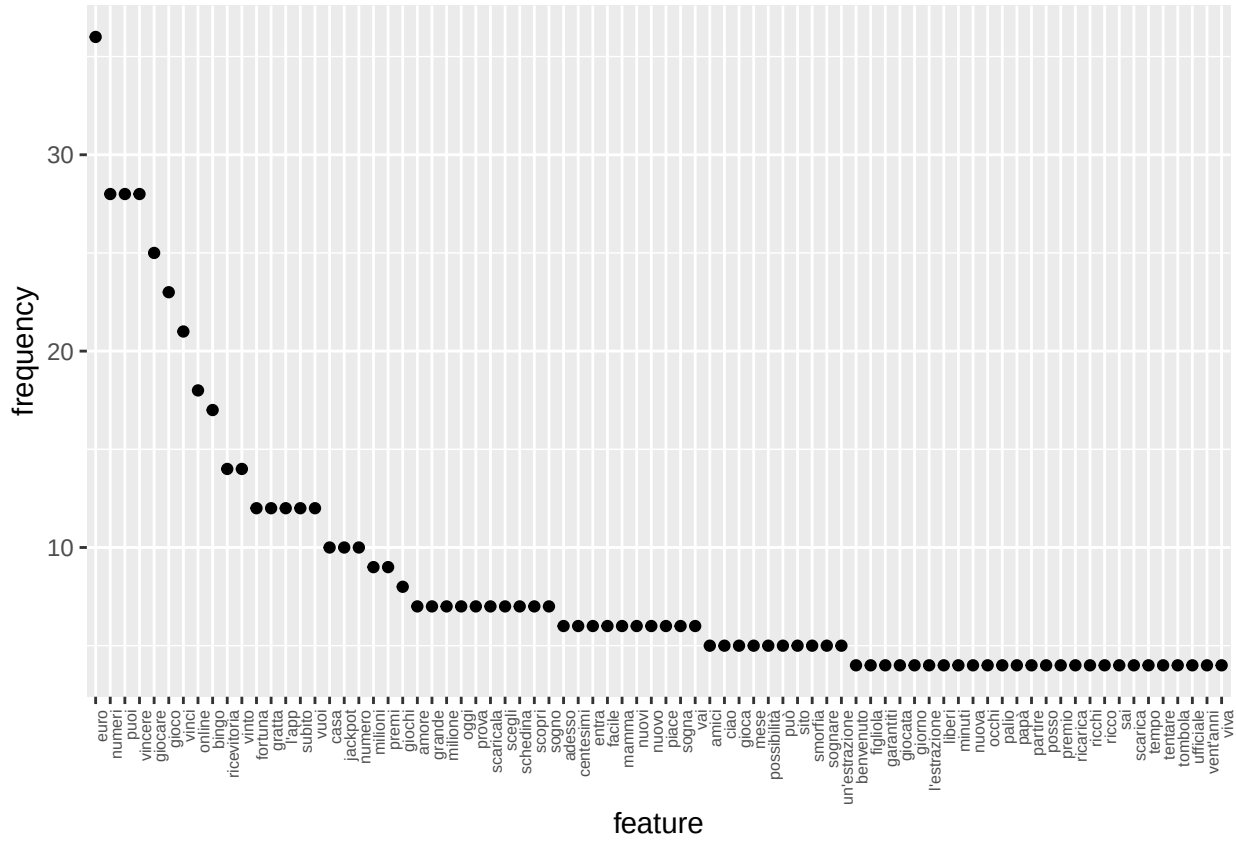


Figure 3: Distribuzione frequenze superiori a 2 testo lotto

```r
tot_poker_ord <- poker4 %>% dfm_weight(scheme = "prop") %>%
  colSums() %>%
  sort(decreasing = TRUE)
barplot(tot_poker_ord[1:10],
        names.arg = names(tot_poker_ord[1:10]),
        col = heat.colors(10), horiz=TRUE,
        main = "poker", xlim = c(0,0.08), las=2)
```

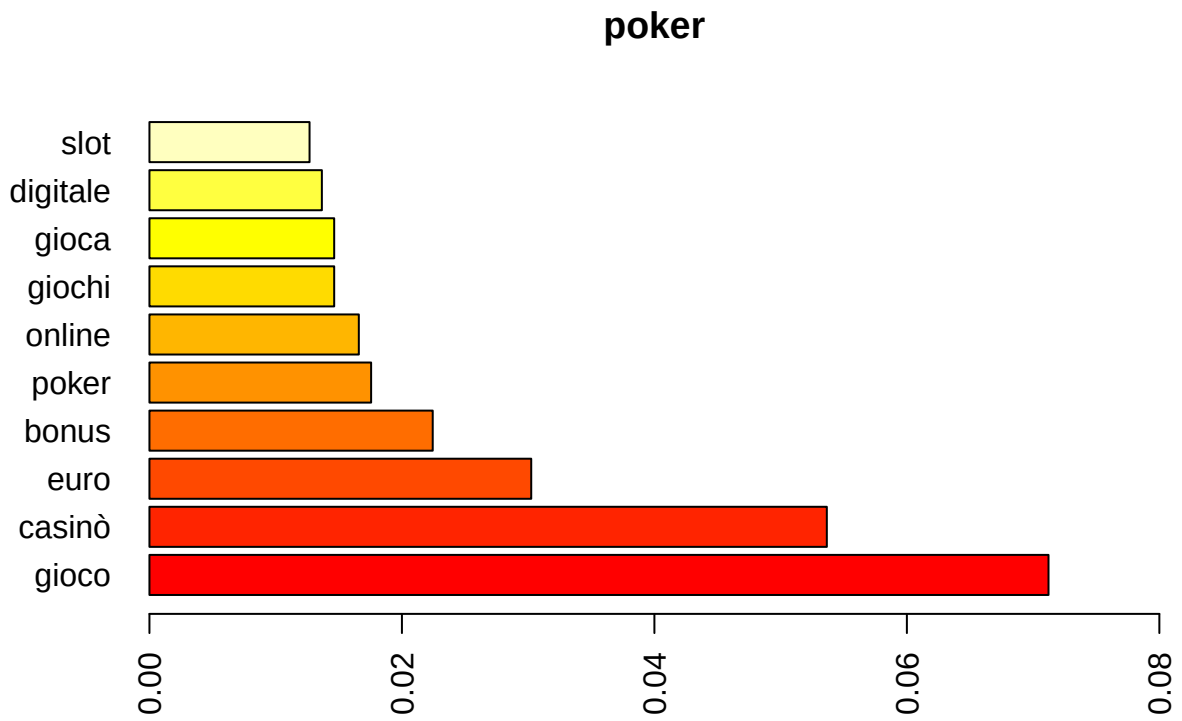


Figure 4: Barplot 10 parole a maggiore frequenza testo poker

```r
tot_sport_ord <- sport4 %>% dfm_weight(scheme = "prop") %>%
  colSums() %>%
  sort(decreasing = TRUE)
barplot(tot_sport_ord[1:10],
        names.arg = names(tot_sport_ord[1:10]),
        col = heat.colors(10), horiz=TRUE,
        main = "sport", xlim = c(0,0.08), las=2)
```

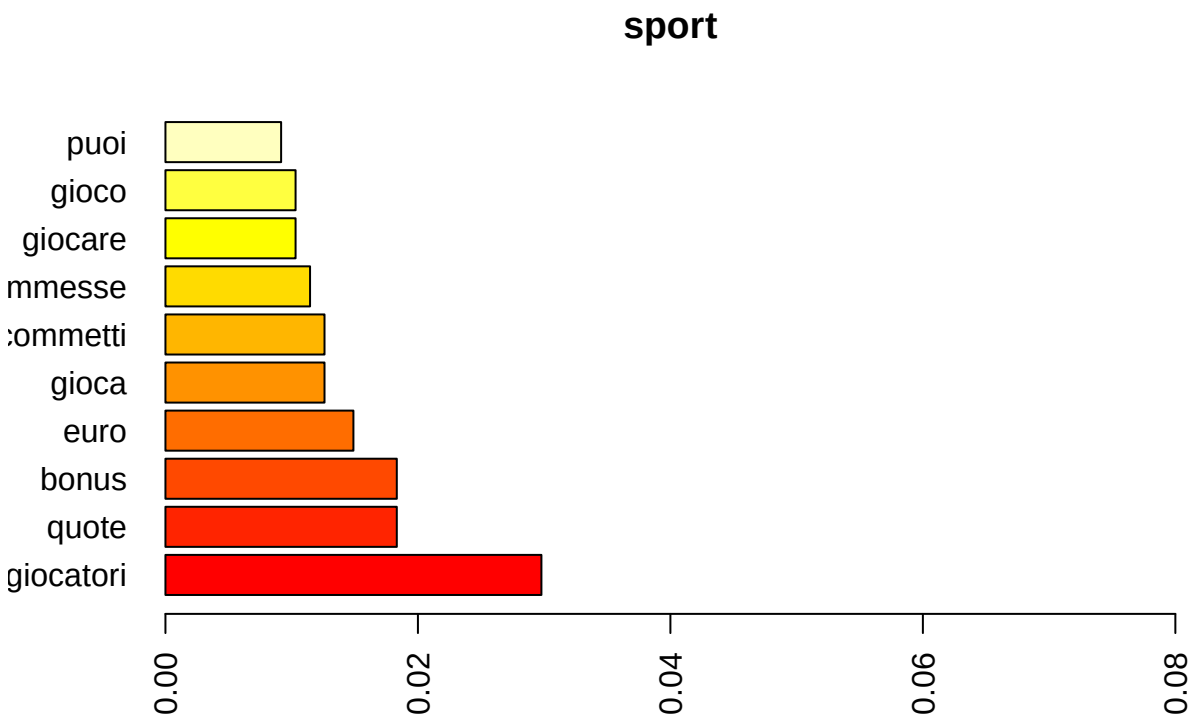


Figure 5: Barplot 10 parole a maggiore frequenza testo sport

```r
tot_lotto_ord <- lotto4 %>% dfm_weight(scheme = "prop") %>%
  colSums() %>%
  sort(decreasing = TRUE)
barplot(tot_lotto_ord[1:10],
        names.arg = names(tot_lotto_ord[1:10]),
        col = heat.colors(10), horiz=TRUE,
        main = "lotto", xlim = c(0,0.08), las=2)
```

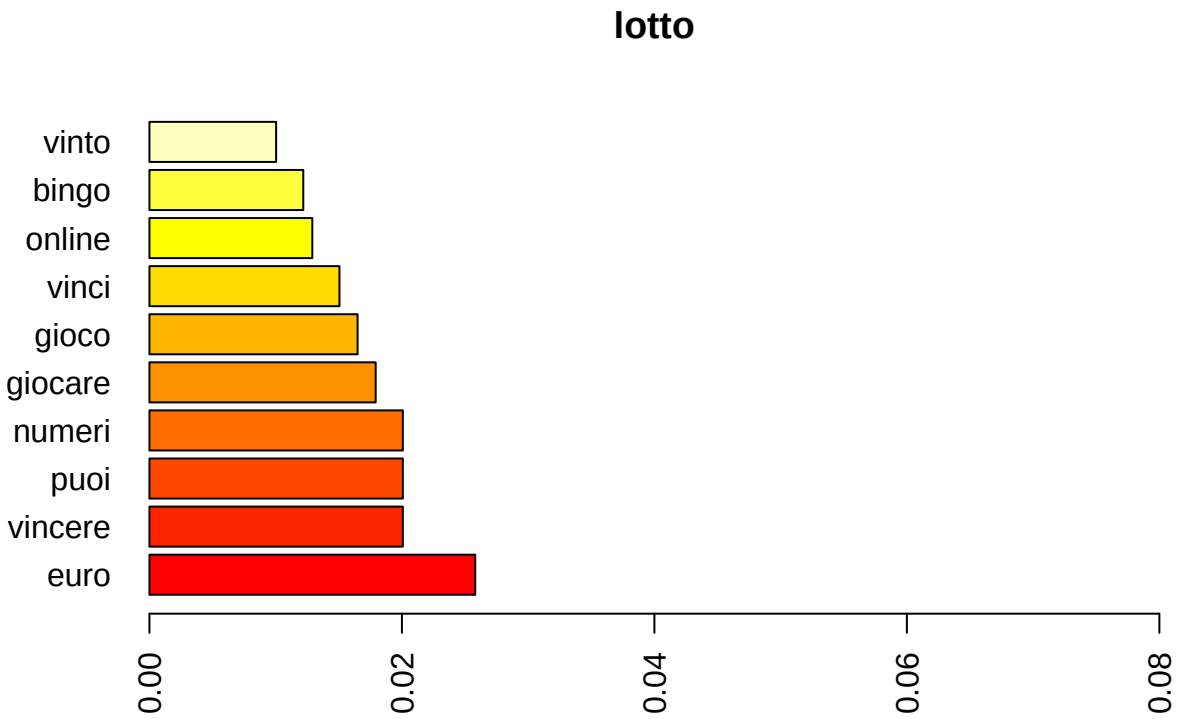


Figure 6: Barplot 10 parole a maggiore frequenza testo lotto

Le distribuzioni appena presentate sono riassumibili attraverso la Legge di Zipf (Zipf 1949). Tale legge prevede che per taluni evento la frequenza sia inversamente proporzionale al proprio rango di appartenenza, disposto in ordine decrescente. Trasportando questa legge al contesto del linguaggio naturale, risulta che tendenzialmente nei *corpora* di qualsiasi tipo le parole appartenenti al primo rango presentano una frequenza doppia rispetto a quelle del secondo e tripla del terzo.

```
Zipf_plot(poker4)
```

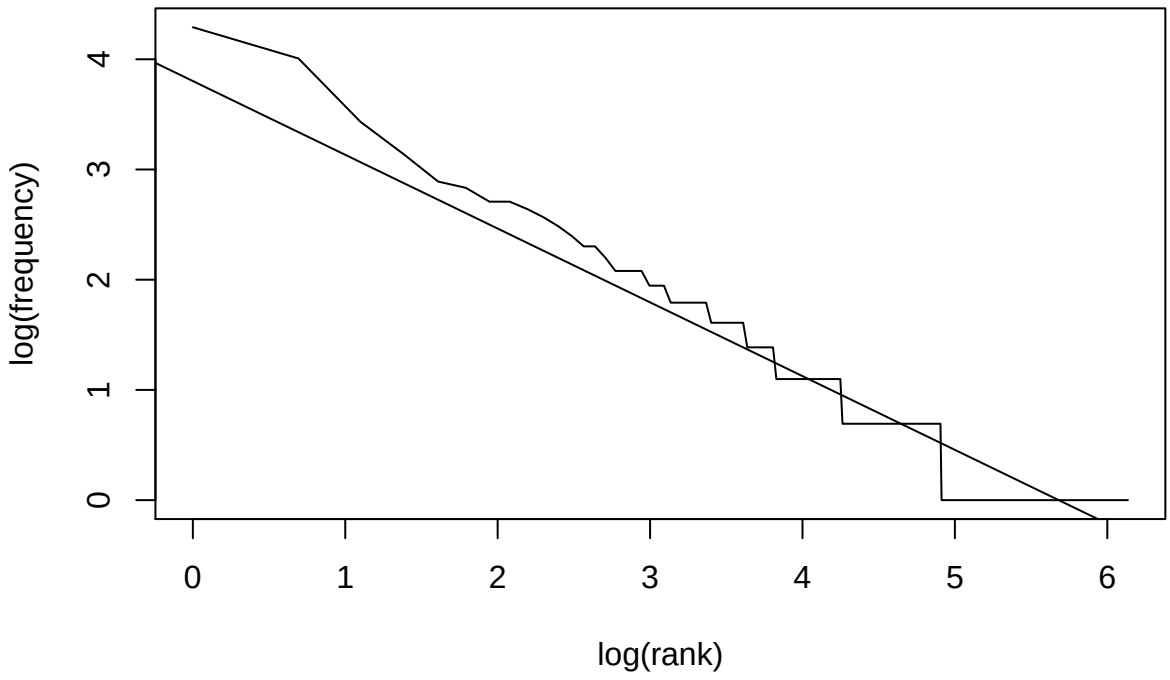


Figure 7: Istogramma Legge di Zipf testo poker

```
(Intercept)           x
  3.8025112  -0.6693043
```

```
Zipf_plot(sport4)
```

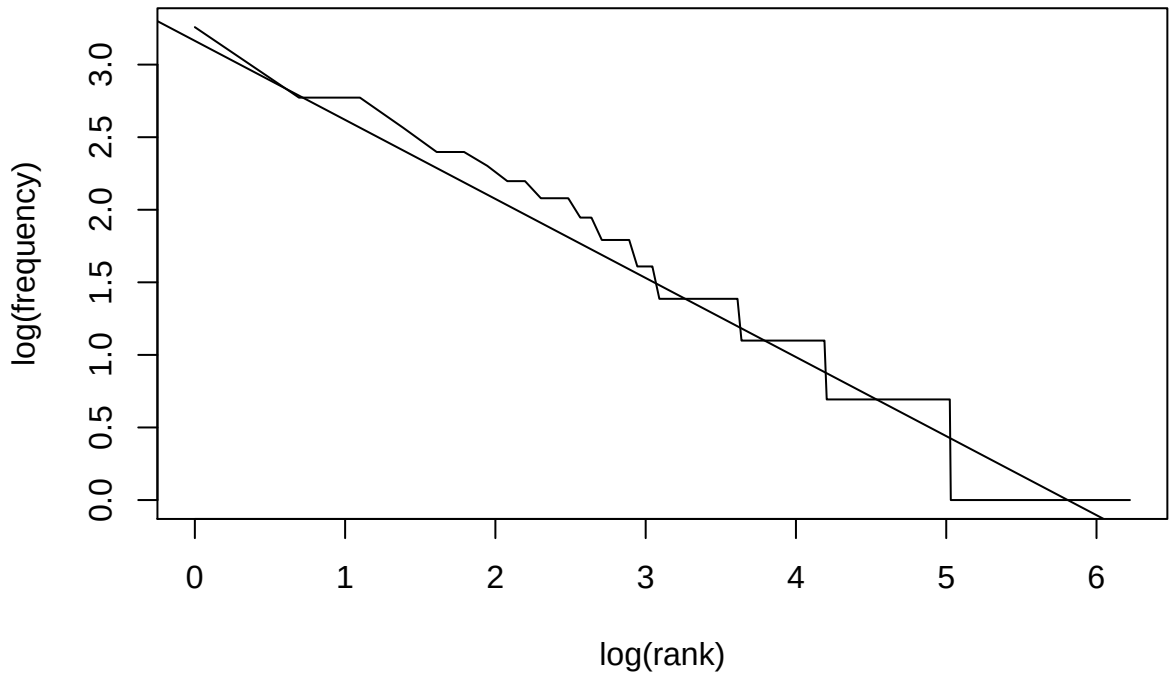


Figure 8: Istogramma Legge di Zipf testo sport

```
(Intercept)           x
  3.1638289  -0.5446959
```

```
Zipf_plot(lotto4)
```

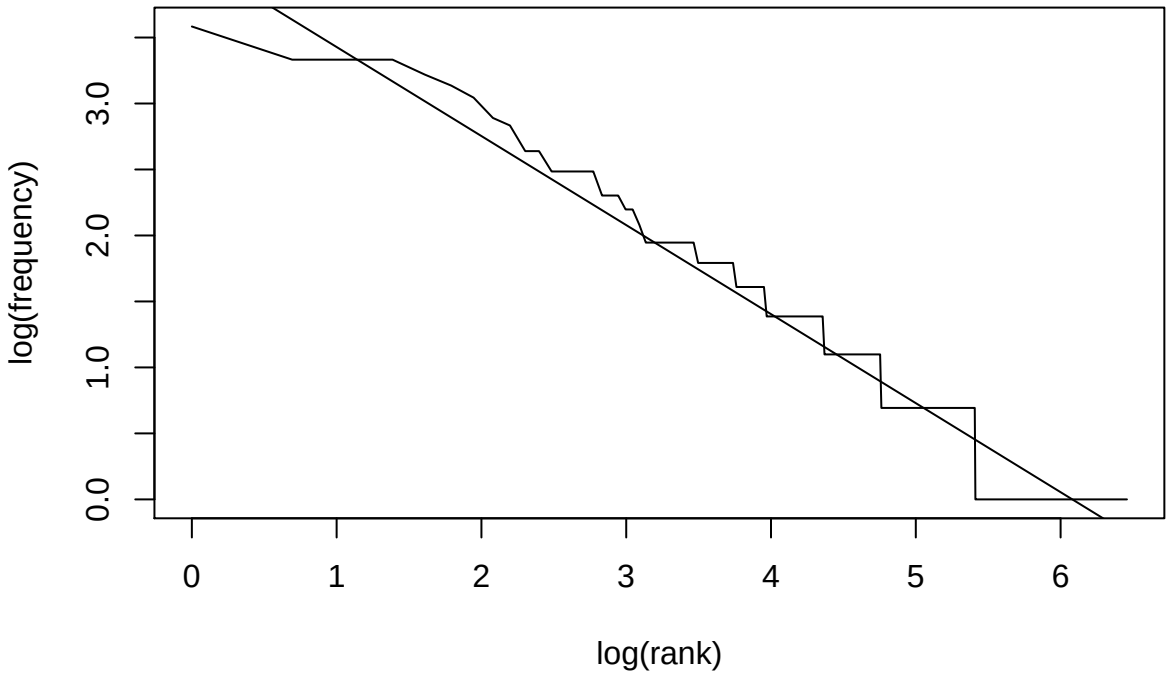


Figure 9: Istogramma Legge di Zipf testo lotto

```
(Intercept)           x
  4.1025826  -0.6746554
```

La Legge di Zipf è utilizzata spesso per individuare i termini che risultano avere un peso rilevante nel significato globale del *corpus*. Generalmente, utilizzando testi non ripuliti, le parole più importanti si collocano in prossimità del centro della retta, in quanto nell'estremità superiore si posizionano quei termini ad altissima frequenza (articoli, congiunzioni) che fungono da collante per creare una sintassi fluida e di senso compiuto; all'estremità inferiore risiedono i termini a occorrenza bassa o unica, per cui di minor rilievo. Nel caso del grafici precedentemente mostrati, la legge è stata applicata ad una matrice `DFM` già ripulita dalle stopwords, per cui i termini contenuti nel primo rango di fatto sono i termini a maggiore rilevanza.

3.3.5 Wordcloud

Le *worcloud* sono i grafici più emblematici per l'NLP, in quanto consentono a colpo d'occhio di avere una comprensione intuitiva ed immediata delle parole chiave appartenenti ad un *corpus*. Nondimeno la loro grafica

è decisamente accattivante e d'impatto. La caratteristica della *wordcloud* è quella di presentare il volume delle parole proporzionato alla loro frequenza.

```
set.seed(100)
textplot_wordcloud(poker4,
                   min_count = 1,
                   random_order = FALSE,
                   rotation = .25,
                   color = c("deepskyblue","dodgerblue3","darkblue"))
```

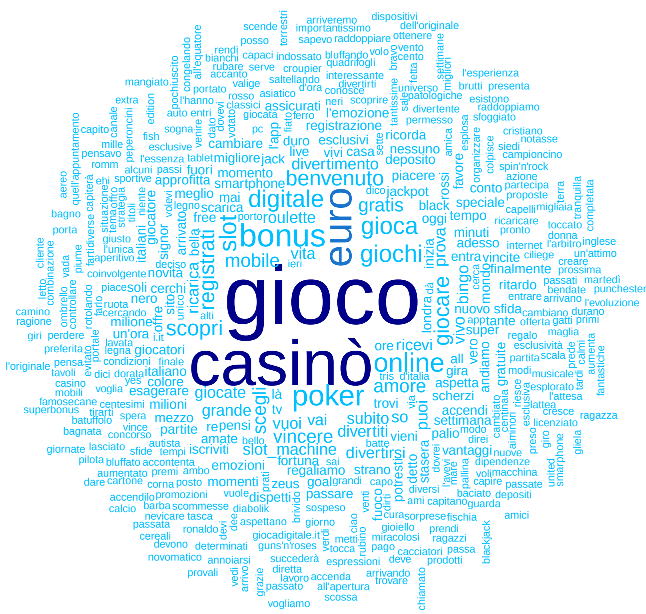


Figure 10: Worcloud poker

```
set.seed(100)
textplot_wordcloud(sport4,
                   min_count = 2,
                   random_order = FALSE,
                   rotation = .25,
                   color = c("deepskyblue","dodgerblue3","darkblue"))
```

Figure 11: Wordcloud sport

```r
set.seed(100)
textplot_wordcloud(lotto4,
                   min_count = 2,
                   random_order = FALSE,
                   rotation = .25,
                   color = c("deepskyblue","dodgerblue3","darkblue"))
```

Figure 12: Wordcloud lotto

```r
set.seed(100)
textplot_wordcloud(azzardo4,
                   min_count = 3,
                   random_order = FALSE,
                   rotation = .25,
                   color = rainbow(3),
                   comparison = TRUE)
```

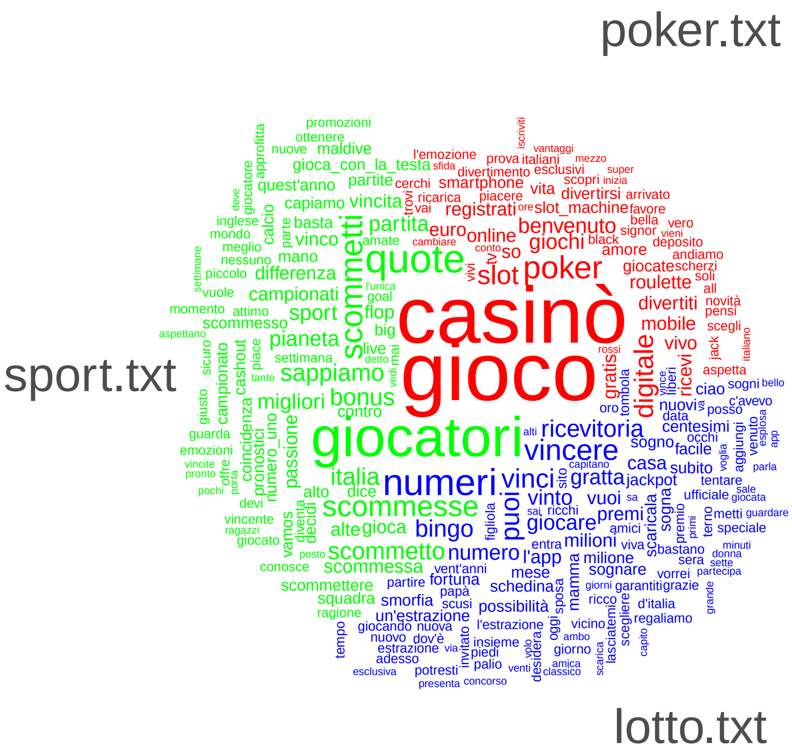


Figure 13: Wordcloud comprensiva di tutti i corpora

Mediante una *wordcloud* comparativa si è in grado di cogliere il differente peso dato alle parole in ciascun copora. Si nota subito che ‘gioco’ e ‘casinò’ sono termini pesantemente calcati negli spot dei giochi in stile casinò rispetto alle altre parole utilizzate nello stesso contesto. Si può ipotizzare che il principale obiettivo promozionale sia quello di creare nel consumatore una ritenzione mnestica dei casinò online. I restanti due corpora invece non presentano dei termini altrettanto marcati, indice di una stesura dei messaggi promozionali leggermente più raffinata e articolata.

3.3.6 Keyness

Mediante la funzione `textstat_keyness` si provvede a visualizzare i termini di maggior rilevanza in ciascun testo confrontato con gli altri. Tale funzione compara le frequenze di parole rilevate in un testo target con le frequenze attese di un *corpus* di confronto (tutti e tre i documenti accorpati). La funzione utilizza la statistica chiquadro e restituisce un p-value indicativo della presenza o meno di una differenza significativa tra le frequenze osservate e quelle attese (permette di osservare se determinate parole hanno un utilizzo maggiore o inferiore rispetto al confronto).

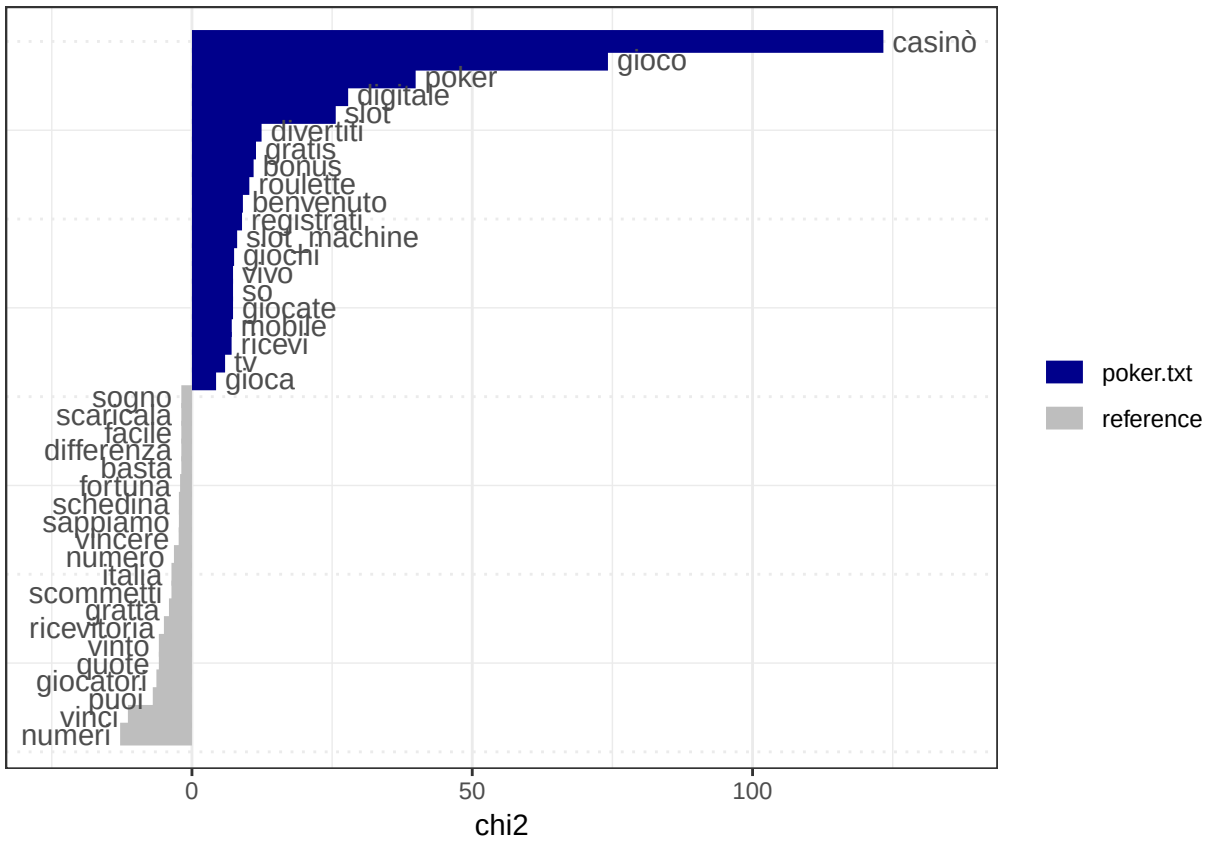
```
library(lubridate)
```

```
Attaching package: 'lubridate'
```

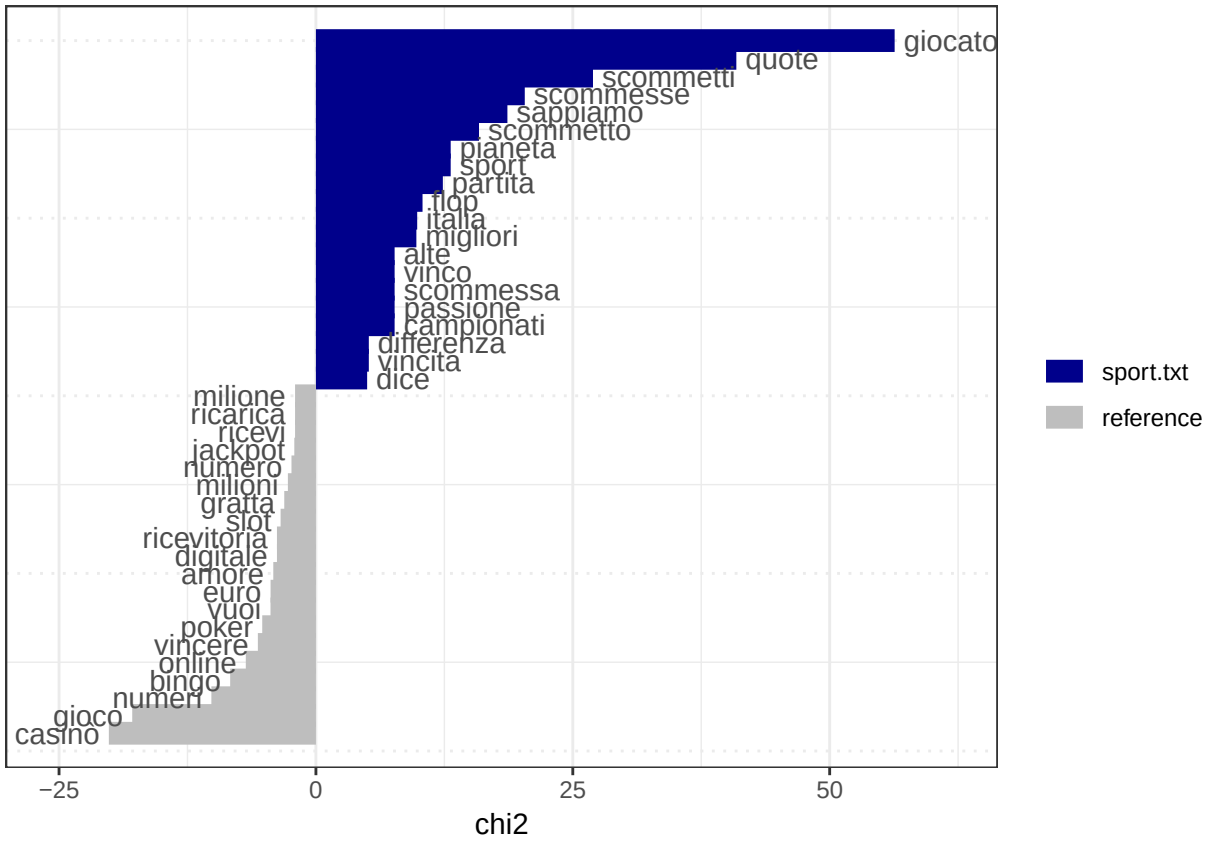
```
The following object is masked from 'package:base':
```

```
    date
```

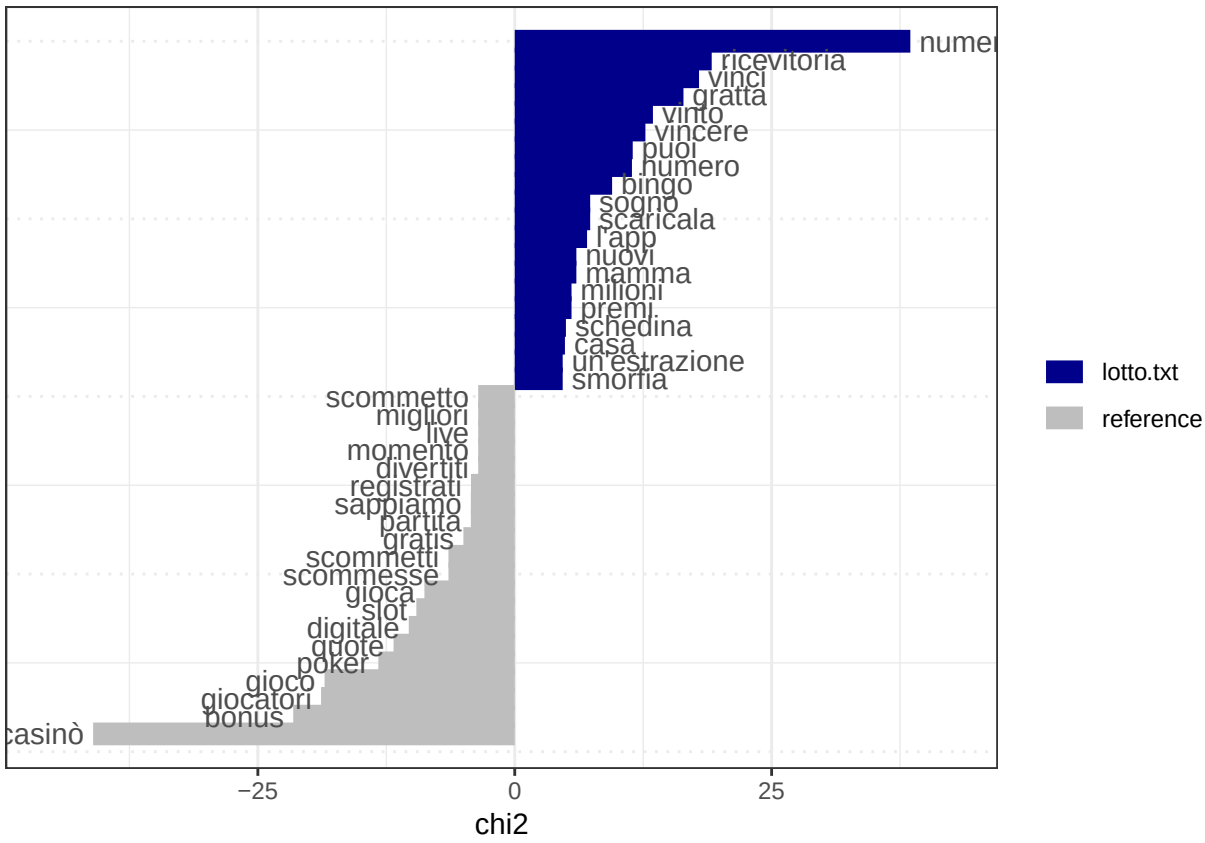
```
tstat_key<-textstat_keyness(azzardo4, target = "poker.txt")
textplot_keyness(tstat_key)
```



```
tstat_key<-textstat_keyness(azzardo4, target = "sport.txt")
textplot_keyness(tstat_key)
```




```
tstat_key<-textstat_keyness(azzardo4, target = "lotto.txt")
textplot_keyness(tstat_key)
```



Questa funzione è utile da sfruttare in abbinata alla *wordcloud* comparativa, perchè permette di esaminare in maniera più robusta il differente utilizzo dei termini.

3.3.7 Semantic network

Tramite la costruzione di una matrice di co-occorrenze è possibile costrire una rete semantica: essa è un potente strumento che permette di visualizzare una rete di concetti relati semanticamente e le connessioni che intercorrono fra loro.

```
featpoker <- names(topfeatures(poker4, 20))
pokerff <- poker4 %>% fcm() %>%
           fcm_select(pattern = featpoker)
set.seed(144)
textplot_network(pokerff,
                 min_freq = 200,
                 edge_alpha = 0.7)
```

```
Registered S3 method overwritten by 'network':
  method            from
```

```
summary.character quanteda
```

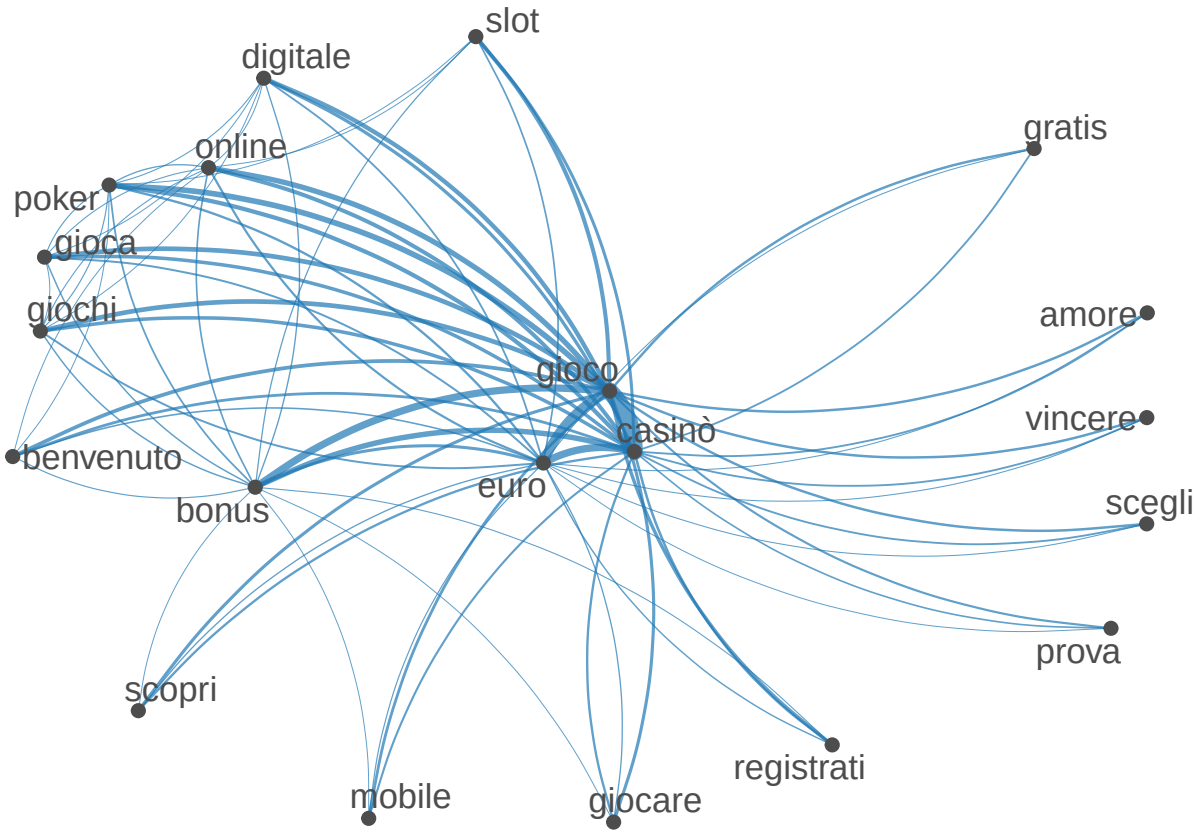


Figure 14: Rete semantica poker

```
set.seed(144)
featsport <- names(topfeatures(sport4, 20))
sport4 %>% fcm() %>%
           fcm_select(pattern = featsport) %>%
           textplot_network(min_freq = 150, edge_alpha = 0.7)
```

```
set.seed(144)
featlotto <- names(topfeatures(lotto4, 20))
lotto4 %>% fcm() %>%
           fcm_select(pattern = featlotto) %>%
           textplot_network(min_freq = 200, edge_alpha = 0.7)
```

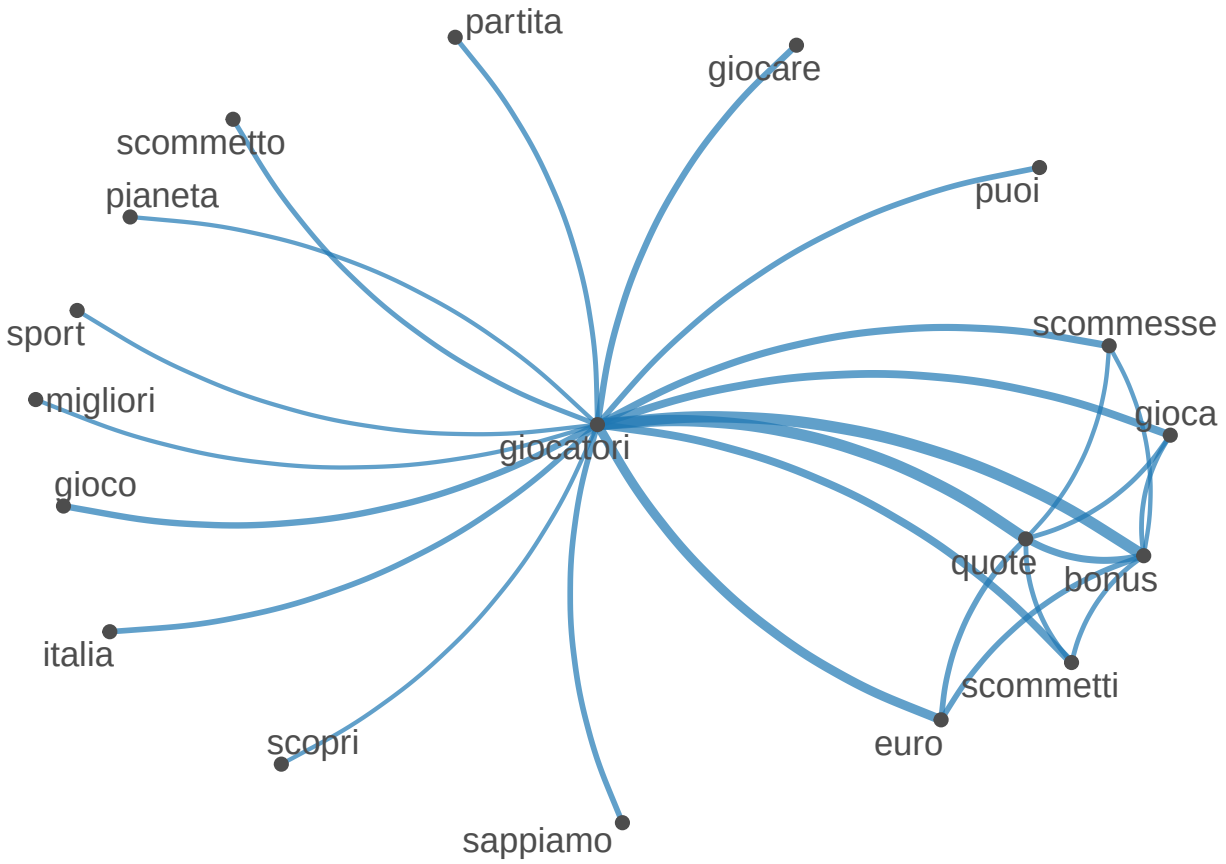


Figure 15: Rete semantica sport

3.3.8 Similarità testuale

Avere tre *corpora* differenti dà la possibilità di poterli confrontare per stabilire quanto essi siano simili tra loro. Per poter svolgere una corretta operazione di paragone è necessario però controllare la lunghezza di ciascuno di essi: l'utilizzo della *cosine similarity* consente di verificare la somiglianza senza che essa sia influenzata dalla magnitudine dei testi. L'utilizzo dell'indice `TF-IDF` (*Type Frequency Inverse Document frequency*) permette di assegnare il giusto peso ai termini che hanno un'importanza maggiore per ciascun singolo documento. Questa funzione opera moltiplicando la frequenza di un termine contenuto in un testo per l'inverso della frequenza dello stesso termine in uno o più testi presi in considerazione. Se un termine presenta un'alta frequenza in un dato testuale ha due possibili spiegazioni: o il termine è effettivamente rilevante, oppure è un termine generalmente ad alta frequenza in qualsiasi testo della stessa lingua (articoli, congiunzioni, etc.). Per discriminare una possibilità dall'altra si moltiplica per l'inverso della frequenza della parola in tutti i documenti: se il valore IDF è alto si può assumere che il termine abbia scarsa importanza.

```
azzardo %>% dfm() %>%
            dfm_tfidf(scheme_tf = "prop", scheme_df = "inverse") %>%
```

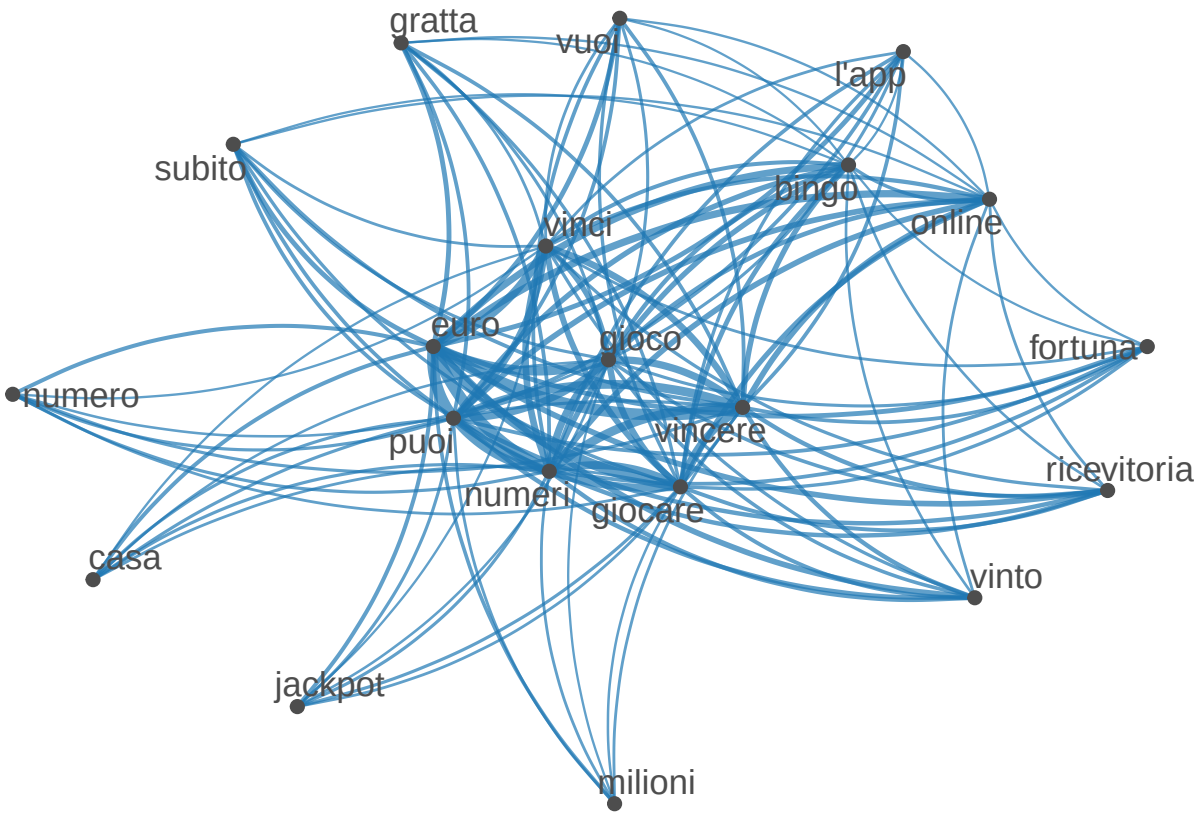


Figure 16: Rete semantica lotto

```
             textstat_simil(method = "cosine")
```

```
            poker.txt   sport.txt
sport.txt 0.006924657
lotto.txt 0.011737571 0.020052514
```

```
tfidf_azzardo<-dfm_tfidf(dfm(azzardo),scheme_tf = "prop", scheme_df = "inverse")
textstat_simil(tfidf_azzardo, method = "cosine")
```

```
            poker.txt   sport.txt
sport.txt 0.006924657
lotto.txt 0.011737571 0.020052514
```

Dalla formula si desume che i testi che presentano maggior somiglianza sono `lotto.txt` e `sport.txt`. Avendo desunto questa informazione, è ragionevole indagare sia gli aspetti di somiglianza che le differenze. Per quanto riguarda la somoglianza è più che sufficiente osservare la distribuzione di frequenza dei termini proporzionata alla lunghezza dei testi.

```
splotto <- sport1 + lotto1
sharedwords_splotto <- splotto  %>%
                       tokens(remove_numbers=TRUE,
                              remove_punct=TRUE,
                              remove_symbols=TRUE) %>%
                       tokens_tolower() %>%
                       tokens_select(stopwords1,
                                     selection = "remove",
                                     padding = FALSE,
                                     verbose = TRUE) %>%
                       dfm(stem = FALSE) %>%
                       dfm_weight(scheme = "boolean") %>%
                       dfm_trim(min_termfreq = 2)
```

```
removed 342 features
```

```
splotto_freq <- splotto %>%
                dfm() %>%
                dfm_select(pattern = sharedwords_splotto, selection = "keep") %>%
                dfm_weight(scheme = "prop") %>%
                textstat_frequency(n = 50)
splotto_freq$feature <- with(splotto_freq, reorder(feature, -frequency))
ggplot(splotto_freq, aes(x = feature, y = frequency)) +
       geom_point() +
       labs(x = "Features", y = "Term frequency") +
       theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

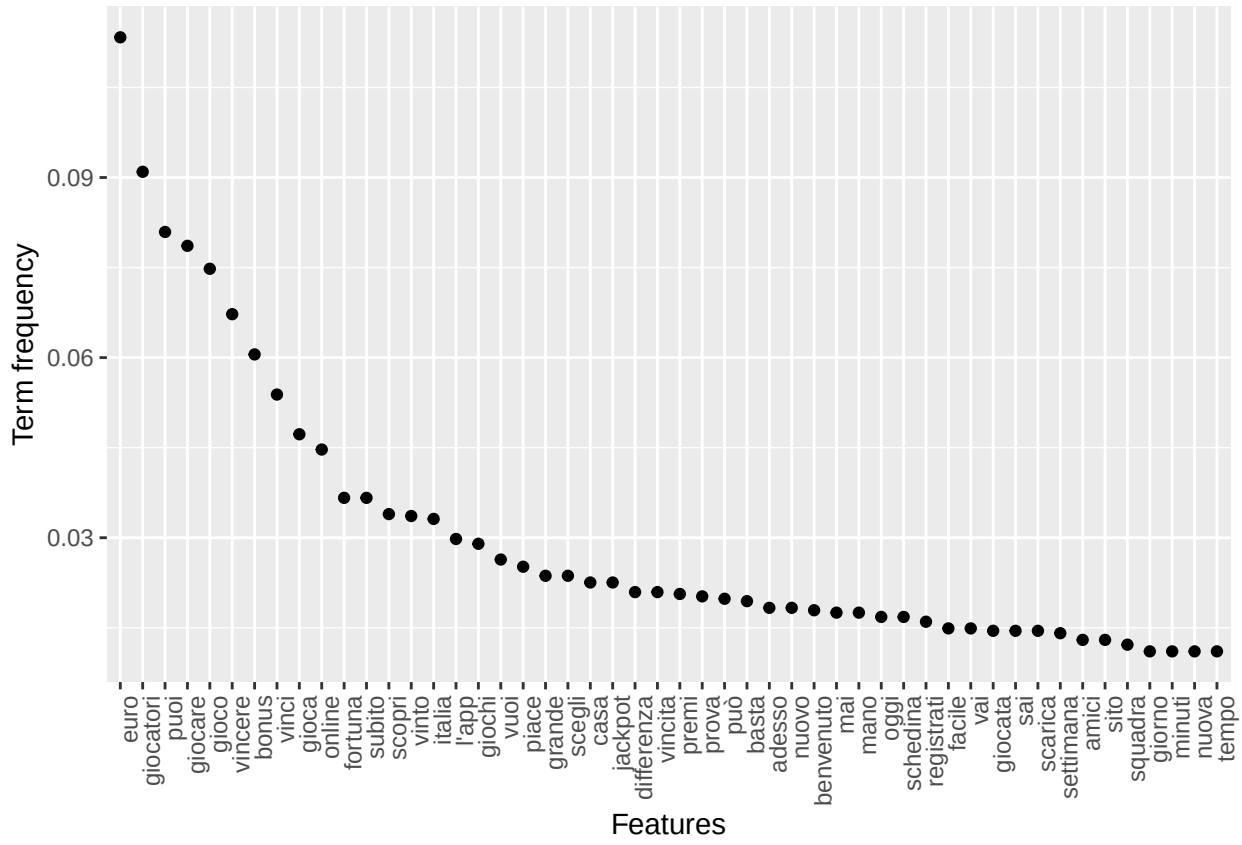


Figure 17: Distribuzione decrescente delle frequenze del vocabolario in comune sport - lotto

In merito alle differenze invece si è stabilito di applicare un modello di *Latent Dirichlet Allocation* (Blei et al. 2003) sulle parole non condivise per estrapolare una serie di topic latenti.

```
library(topicmodels)
sport_paroleuniche <- sport4 %>% dfm_remove(sharedwords_splotto)


lda_sport_paroleuniche <- convert(sport_paroleuniche,
                                  to = "topicmodels") %>%
                          LDA(k = 3)
get_terms(lda_sport_paroleuniche, 10)
```

```
      Topic 1      Topic 2     Topic 3
 [1,] "quote"      "scommesse" "quote"
 [2,] "scommetti"  "scommetto" "scommesse"
 [3,] "campionati" "partita"   "migliori"
 [4,] "partita"    "sappiamo"  "scommetti"
```

```
 [5,] "pianeta"    "quote"     "sappiamo"
 [6,] "flop"       "scommetti" "sport"
 [7,] "meglio"     "contro"    "vinco"
 [8,] "live"       "dice"      "flop"
 [9,] "passione"   "pianeta"   "vamos"
[10,] "sport"      "partite"   "partite"
```

```
lotto_paroleuniche <- lotto4 %>% dfm_remove(sharedwords_splotto)
lda_lotto_paroleuniche <- convert(lotto_paroleuniche,
                                  to = "topicmodels") %>%
                          LDA(k = 3)
get_terms(lda_lotto_paroleuniche, 10)
```

```
      Topic 1       Topic 2       Topic 3
 [1,] "ricevitoria" "numeri"      "numeri"
 [2,] "numeri"      "bingo"       "bingo"
 [3,] "entra"       "gratta"      "gratta"
 [4,] "milioni"     "ricevitoria" "premio"
 [5,] "numero"      "scaricala"   "amore"
 [6,] "sogno"       "milioni"     "garantiti"
 [7,] "mamma"       "centesimi"   "nuovi"
 [8,] "sognare"     "numero"      "ricchi"
 [9,] "gratta"      "amore"       "numero"
[10,] "vent'anni"   "sogno"       "ricco"
```

I topic ricavati non risultano essere molto informativi. L'unica nozione che si ricava riguarda il fatto che il testo `lotto` preferisce calcare sui termini relativi ai numeri mentre il testo `sport` predilige i termini relativi alle scommesse. Non appaiono differenze vistose nel confronto intra topic dello stesso documento di appartenenza.

3.3.9 Hierarchical Clustering

Uno step essenziale per lo studio statistico dei testi è l'analisi dei cluster. Tramite l'utilizzo delle distanze interne al testo tra una parola e l'altra si ricavano i gruppi di parole che occorrono insieme a maggior frequenza. In tal modo è possibile ricavare dei cluster distintivi di parole relative ad un testo e osservarne la loro gerarchia.

```
library(cluster)
library(dendextend)
dend_poker <- poker4 %>% dfm_trim(min_termfreq = 3) %>%
```

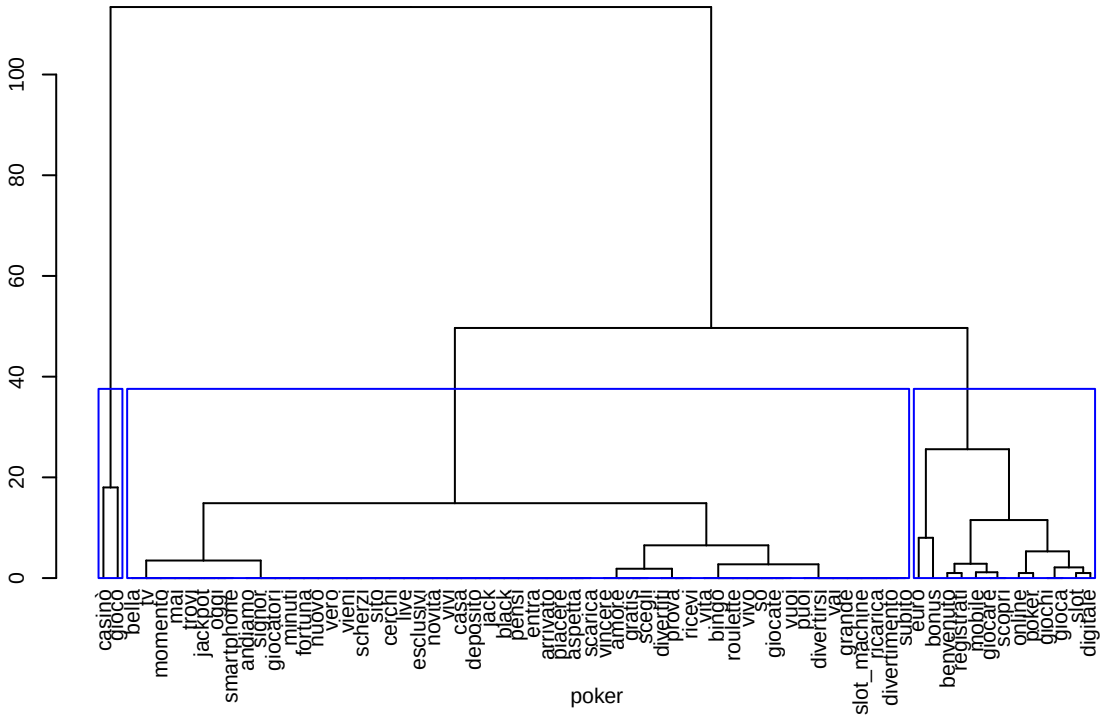


Figure 18: Dendrogramma poker

```
              t() %>%
              dist(method = "euclidean")  %>%
              hclust(method = "ward.D2") %>%
              as.dendrogram()
par(cex = 0.7)
plot(dend_poker, xlab = "poker")
rect.dendrogram(dend_poker, k = 3, border = "blue", xpd = FALSE, lower_rect = 0)
```

```
dend_sport <- sport4 %>% dfm_trim(min_termfreq = 3) %>%
              t() %>%
              dist(method = "euclidean")  %>%
              hclust(method = "ward.D2") %>%
              as.dendrogram()
par(cex = 0.7)
plot(dend_sport, xlab = "sport")
```

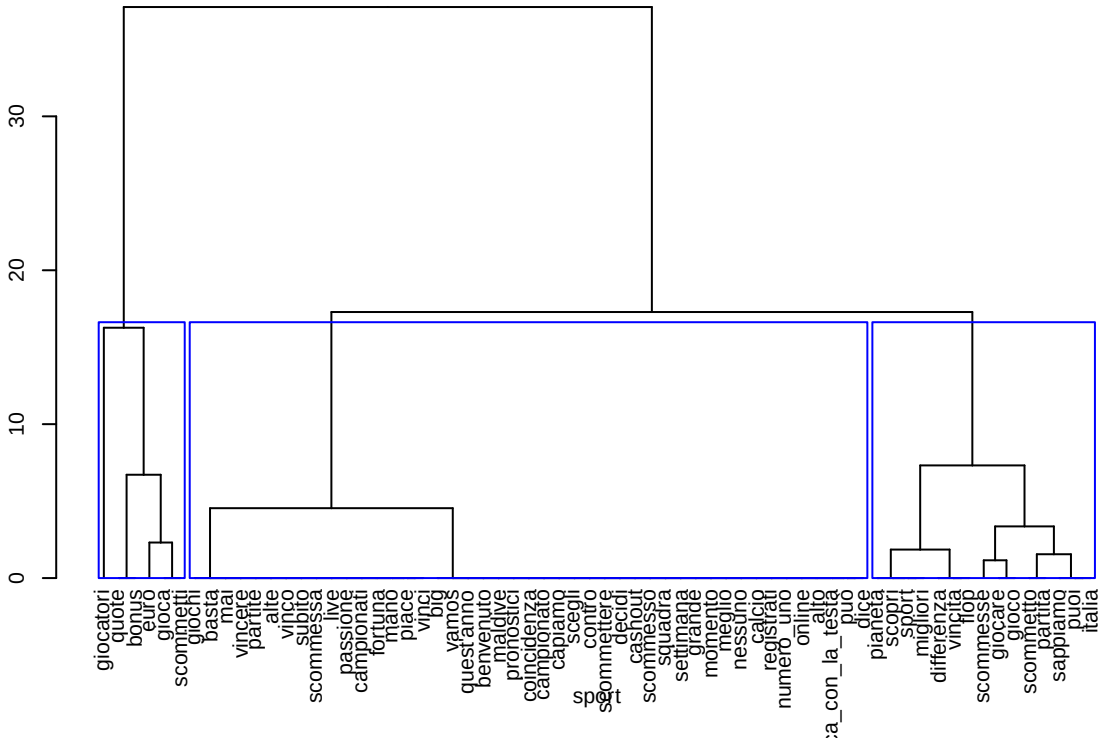


Figure 19: Dendrogramma sport

```
rect.dendrogram(dend_sport, k = 3, border = "blue", xpd = FALSE, lower_rect = 0)
```

```
dend_lotto <- lotto4 %>% dfm_trim(min_termfreq = 3) %>%
              t() %>%
              dist(method = "euclidean")  %>%
              hclust(method = "ward.D2") %>%
              as.dendrogram()
par(cex = 0.5)
plot(dend_lotto, xlab = "lotto")
rect.dendrogram(dend_lotto, k = 5, border = "blue", xpd = FALSE, lower_rect = 0)
```

Una volta eseguiti i grafici singoli di ciascun *corpus* si opera un confronto tra i dendrogrammi dei testi che si sono presentati più simili tra loro (`sport.txt` e `lotto.txt`). In questo modo si è in grado di confrontare le differenze nell'utilizzo dei termini e nella rete gerarchica.

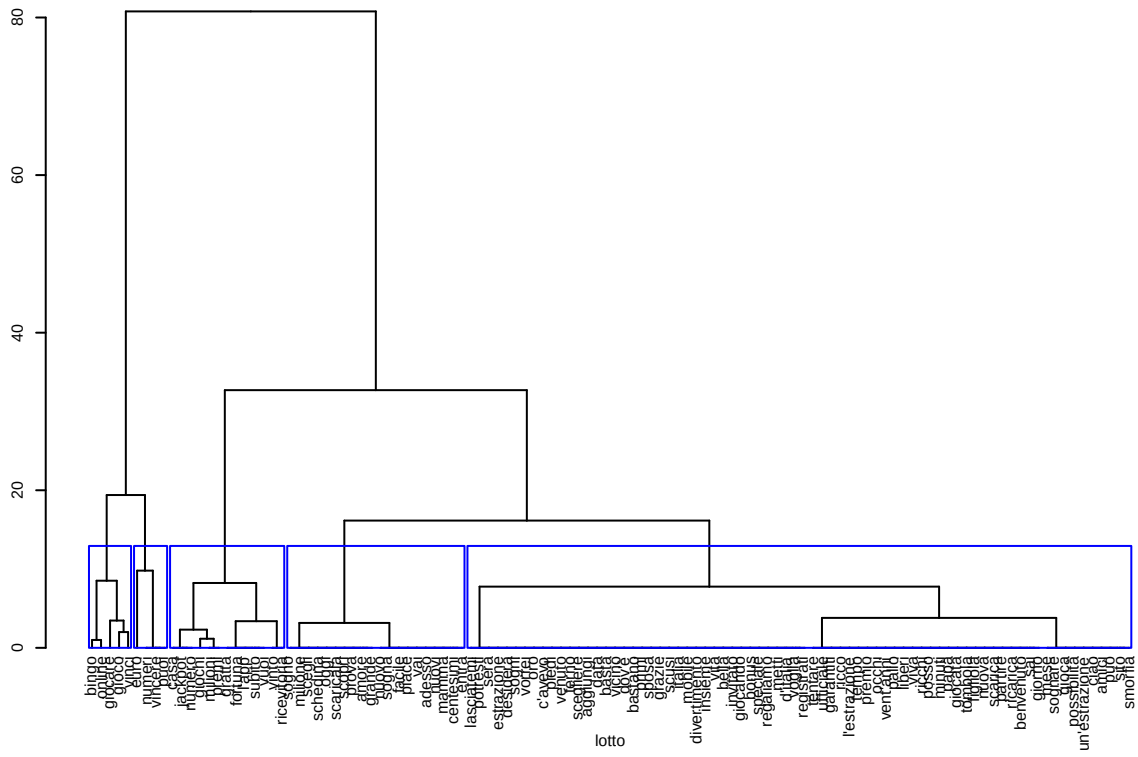


Figure 20: Dendrogramma lotto

```
boosplotto <- rbind(lotto4, sport4) %>% dfm_weight(scheme = "boolean")
commonwords_splotto <- boosplotto %>% dfm_trim(min_termfreq = 2) %>%
                       featnames() %>%
                       tokens()
hc_lotto_common_words <- lotto4 %>% dfm_select(commonwords_splotto) %>%
                         dfm_trim(min_termfreq = 3) %>%
                         t() %>%
                         dist(method = "euclidean") %>%
                         hclust(method = "ward.D2")
hc_sport_common_words <- sport4 %>% dfm_select(commonwords_splotto) %>%
                         dfm_trim(min_termfreq = 3) %>%
                         t() %>%
                         dist(method = "euclidean") %>%
                         hclust(method = "ward.D2")
```

```
tanglegram(hc_lotto_common_words,
           hc_sport_common_words,
           main_left = "lotto",
           main_right = "sport",
           lwd = 1,
           columns_width = c(0.8,0.3,0.5),
           type = "r",
           highlight_distinct_edges = FALSE,
           highlight_branches_lwd = FALSE,
           cex_main = 2)
```

Da cui si ricavano alcuni indici.

- Entanglement: misura che varia tra 0 e 1 e indica il grado di intreccio nella disposizione delle parole. A 0 corrisponde una disposizione perfettamente identica delle parole tra i due testi. A 1 corrisponde una diposizione perfettamente inversa.

```
dend_lotto <- lotto4 %>% dfm_trim(min_termfreq = 1) %>%
              t() %>%
              dist(method = "euclidean") %>%
              hclust(method = "ward.D2") %>%
              as.dendrogram()
dend_sport <- sport4 %>% dfm_trim(min_termfreq = 1) %>%
```

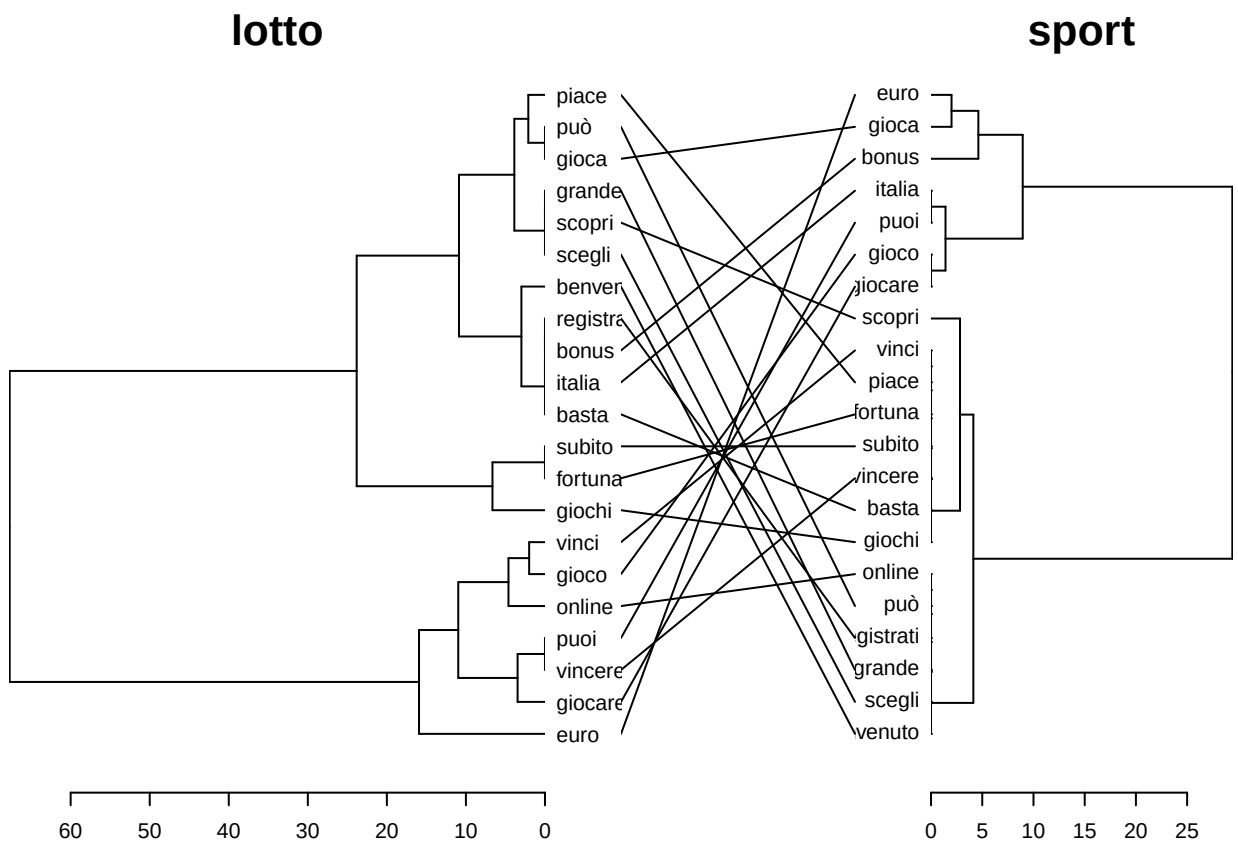


Figure 21: Grafico entanglement lotto - sport

```
              t() %>%
              dist(method = "euclidean") %>%
              hclust(method = "ward.D2") %>%
              as.dendrogram()
```

```
entanglement(intersect_trees(dend1 = dend_lotto, dend2 = dend_sport))
```

```
[1] 0.5799524
```

- Correlazione Gamma di Baker (Baker 1974): misura di similarità tra dendogrammi che varia tra -1 e 1. Un valore pari a 0 significa che i due dendrogrammi non sono statisticamente simili.

```
cor_bakers_gamma(intersect_trees(dend1 = dend_lotto, dend2 = dend_sport))
```

```
[1] 0.224922
```

- Correlazione dei nodi in comune: misura che varia tra 0 e 1 indicativa di quanti nodi sono in comune tra un dendrogramma e l'altro (un nodo è la sezione del grafico da cui si dipanano i rami relativi a

ciascun termine). A 0 corrisponde una assenza totale di nodi uguali, a 1 una perfetta corrispondenza.

```
trim_dend_lotto <- lotto4 %>% dfm_trim(min_termfreq = 2) %>%
                   t() %>%
                   dist(method = "euclidean") %>%
                   hclust(method = "ward.D2") %>%
                   as.dendrogram()
trim_dend_sport <- sport4 %>% dfm_trim(min_termfreq = 2) %>%
                   t() %>%
                   dist(method = "euclidean") %>%
                   hclust(method = "ward.D2") %>%
                   as.dendrogram()
```

```
cor_common_nodes(dend1 = trim_dend_lotto, dend2 = trim_dend_sport)
```

3.3.10 Universal Part of Speech Tagging (UPOS)

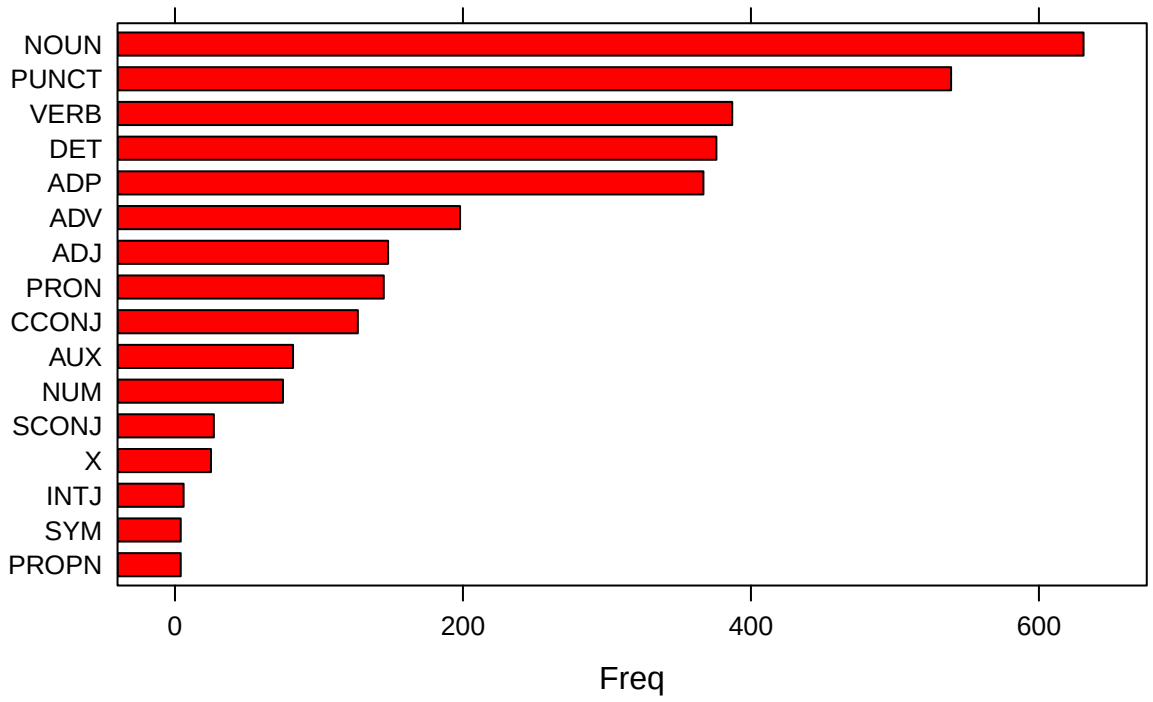
Ad ogni corpus è stata effettuata un'operazione di *tagging* per poter effettuare un'analisi grammaticale delle parole. L'operazione è stata svolta mediante l'utilizzo di un particolare vocabolario d'italiano integrato al software di analisi. Dal momento che la funzione di riconoscimento è sensibile al contesto in cui sono inserite le parole, il lavoro è svolto sui corpora ripuliti solamente delle avvertenze sui rischi del gioco d'azzardo.

```
library(udpipe)
library(lattice)
library(tm)
poker1 <- poker %>% tm::VectorSource() %>%
          tm::Corpus() %>%
          tm::tm_map(removeWords, c("minori", "vietato",
                                    "maggiorenni","patologica",
                                    "può","causare", "dipendenza")) %>%
          quanteda::corpus(.$content)  %>%
          .$documents %>%
          .$texts %>%
          quanteda::char_tolower()
kwic(poker1, pattern = "dipendenza")
```

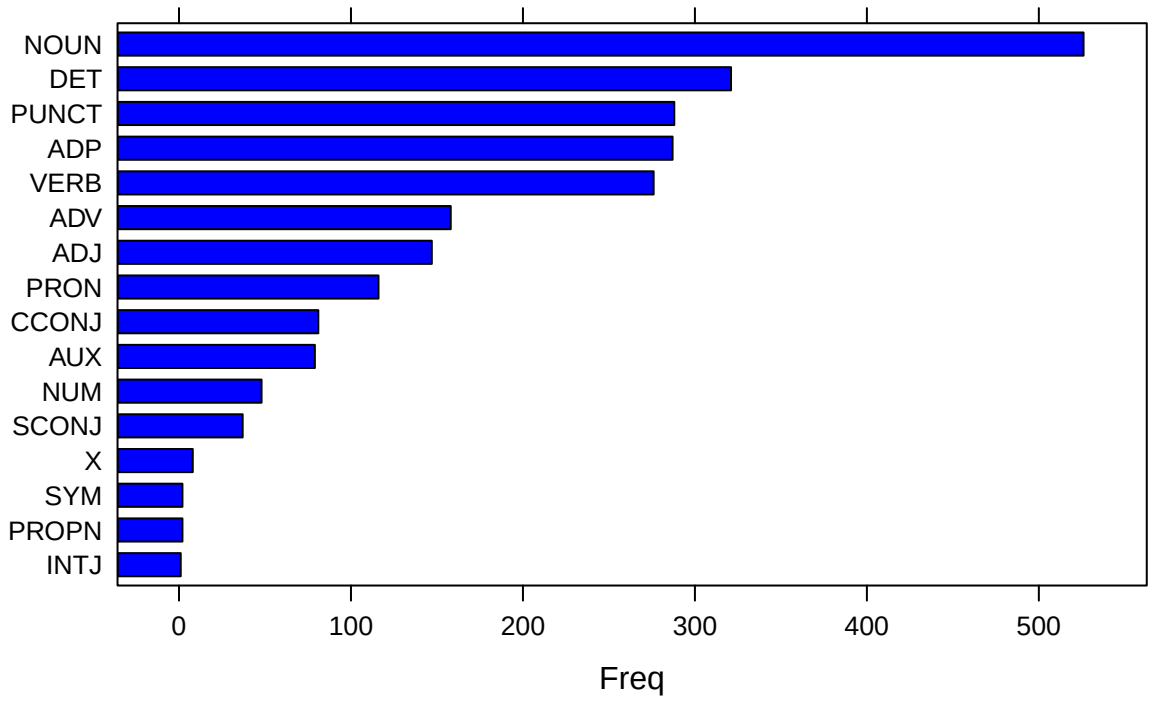
```
kwic object with 0 rows
```

```
udmodel_italian <- udpipe_load_model(
  "/home/niccolo/Scrivania/TESI/tesi/txt/italian-isdt-ud-2.3-181115.udpipe")
dataframe_poker <- poker1 %>% udpipe::udpipe_annotate(udmodel_italian, .) %>%
                              data.frame()
dataframe_sport <- sport1$documents$text %>% quanteda::char_tolower() %>%
                   udpipe::udpipe_annotate(udmodel_italian, .) %>%
                   data.frame()
dataframe_lotto <- lotto1$documents$text %>% quanteda::char_tolower() %>%
                   udpipe::udpipe_annotate(udmodel_italian, .) %>%
                   data.frame()
stats_poker<-txt_freq(dataframe_poker$upos)
stats_poker$key <- factor(stats_poker$key, levels = rev(stats_poker$key))
stats_sport<-txt_freq(dataframe_sport$upos)
stats_sport$key <- factor(stats_sport$key, levels = rev(stats_sport$key))
stats_lotto<-txt_freq(dataframe_lotto$upos)
stats_lotto$key <- factor(stats_lotto$key, levels = rev(stats_lotto$key))
```

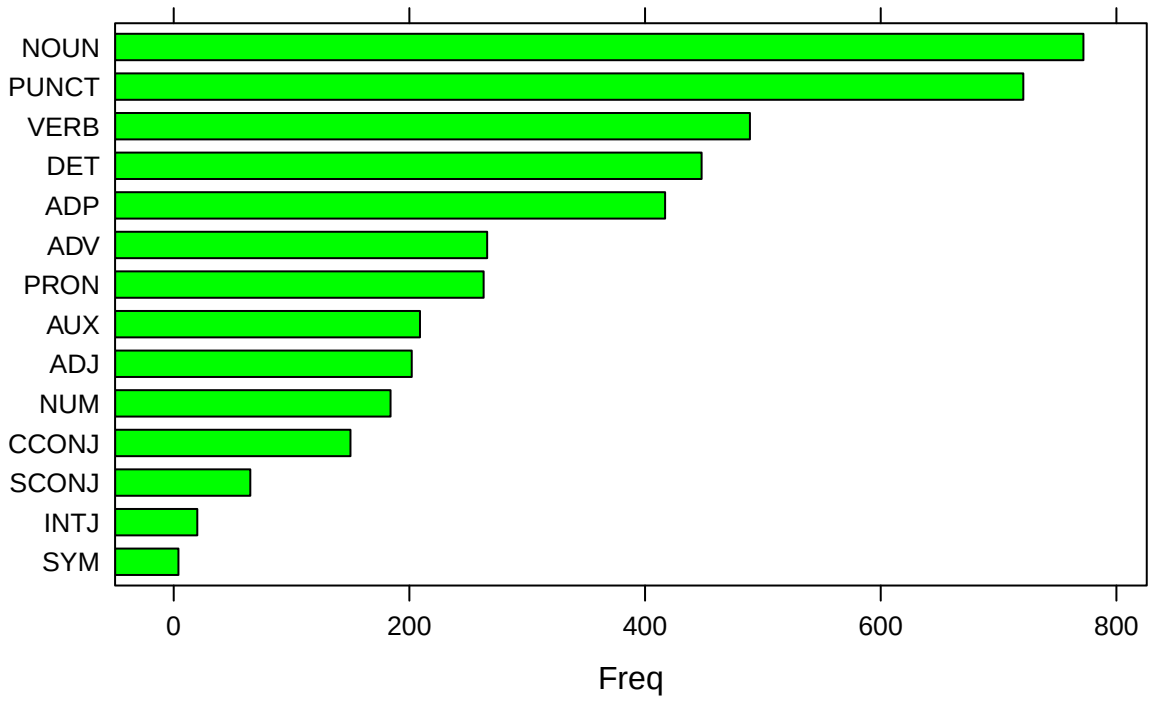
```
barchart(key ~ freq, data = stats_poker, col = "red",
             main = "UPOS Poker\n frequency of occurrence",
             xlab = "Freq")
```




```
barchart(key ~ freq, data = stats_sport, col = "blue",
              main = "UPOS Sport\n frequency of occurrence",
              xlab = "Freq")
```




```
barchart(key ~ freq, data = stats_lotto, col = "green",
             main = "UPOS Lotto\n frequency of occurrence",
             xlab = "Freq")
```




```
stats_poker_noun <- subset(dataframe_poker, upos %in% c("NOUN"))
stats_poker_noun <- txt_freq(stats_poker_noun$token)
sum(stats_poker_noun$freq)
```

```
[1] 631
```

```
stats_poker_noun$key <- factor(stats_poker_noun$key,
                               levels = rev(stats_poker_noun$key))
barchart(key ~ freq, data = head(stats_poker_noun, 10),col = "red",
         main = "Nomi più frequenti Poker",
         xlab = "Freq")
```

```
stats_poker_adj <- subset(dataframe_poker, upos %in% c("ADJ"))
stats_poker_adj <- txt_freq(stats_poker_adj$token)
stats_poker_adj$key <- factor(stats_poker_adj$key,
                              levels = rev(stats_poker_adj$key))
barchart(key ~ freq, data = head(stats_poker_adj, 10), col = "red",
         main = "Aggettivi più frequenti Poker",
         xlab = "Freq")
```

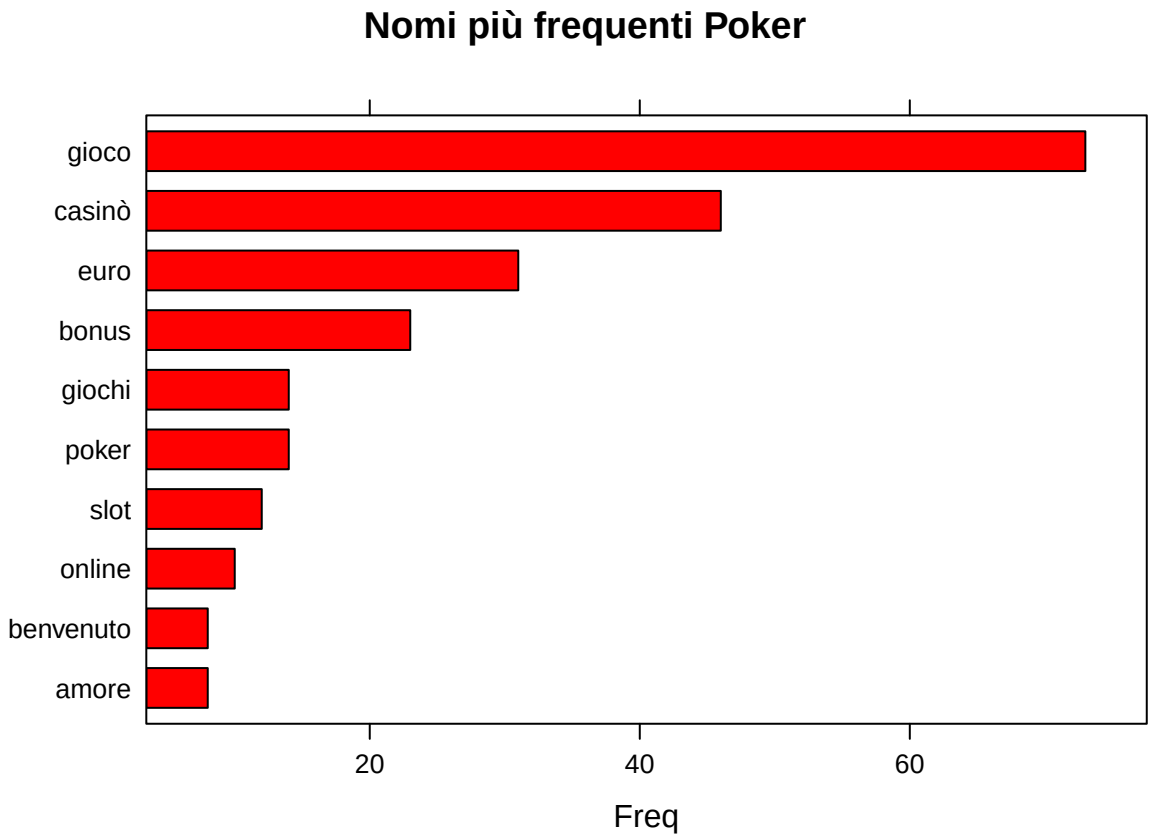


Figure 22: Barplot nomi poker

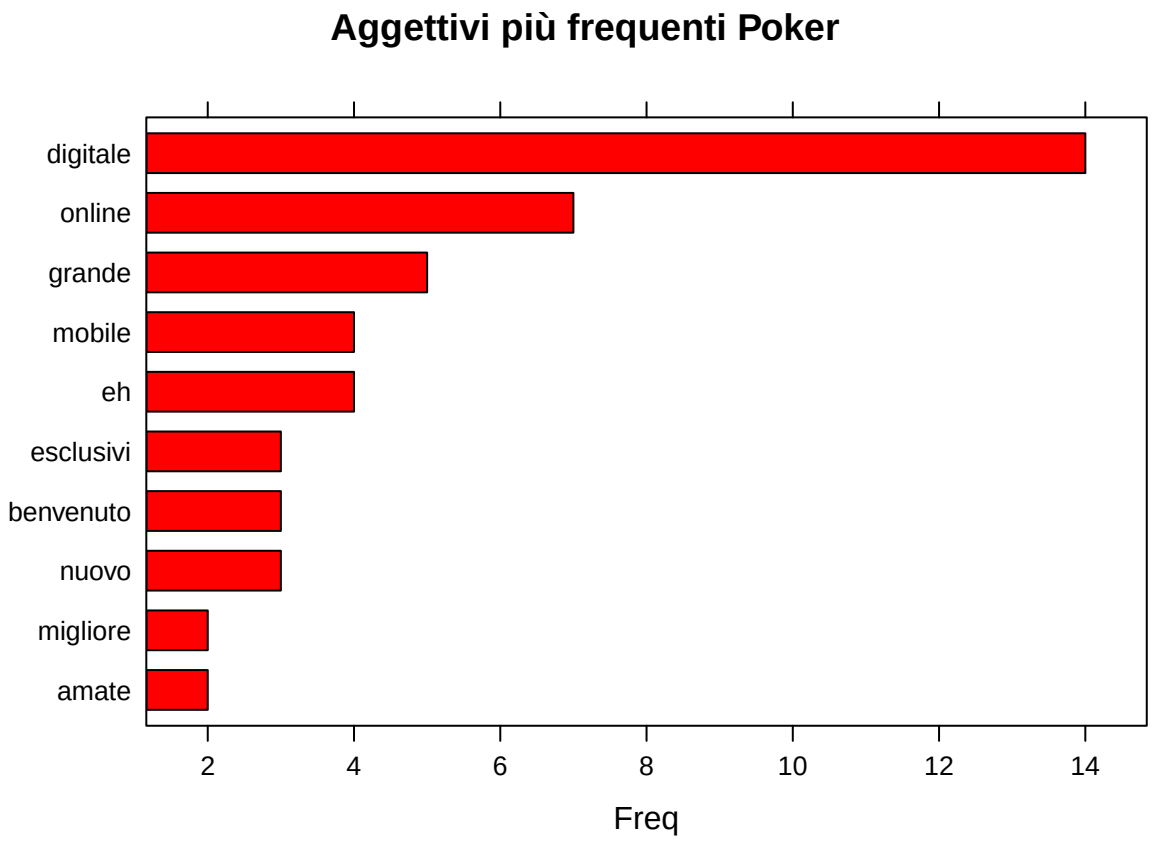


Figure 23: Barplot aggettivi poker

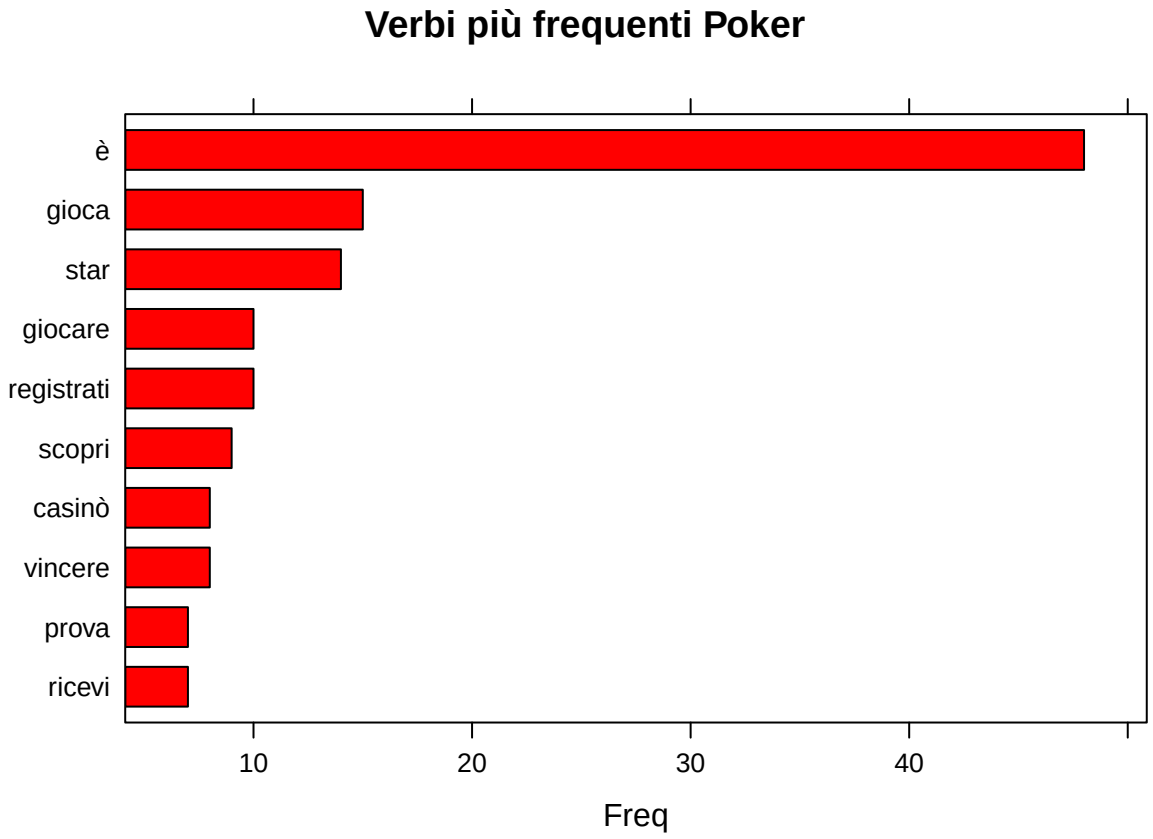


Figure 24: Barplot verbi poker

```
stats_poker_verb <- subset(dataframe_poker, upos %in% c("VERB"))
stats_poker_verb <- txt_freq(stats_poker_verb$token)
stats_poker_verb$key <- factor(stats_poker_verb$key,
                               levels = rev(stats_poker_verb$key))
barchart(key ~ freq, data = head(stats_poker_verb, 10), col = "red",
         main = "Verbi più frequenti Poker",
         xlab = "Freq")
```

```
stats_sport_noun <- subset(dataframe_sport, upos %in% c("NOUN"))
stats_sport_noun <- txt_freq(stats_sport_noun$token)
stats_sport_noun$key <- factor(stats_sport_noun$key,
                               levels = rev(stats_sport_noun$key))
barchart(key ~ freq, data = head(stats_sport_noun, 10),col = "blue",
         main = "Nomi più frequenti Sport",
         xlab = "Freq")
```

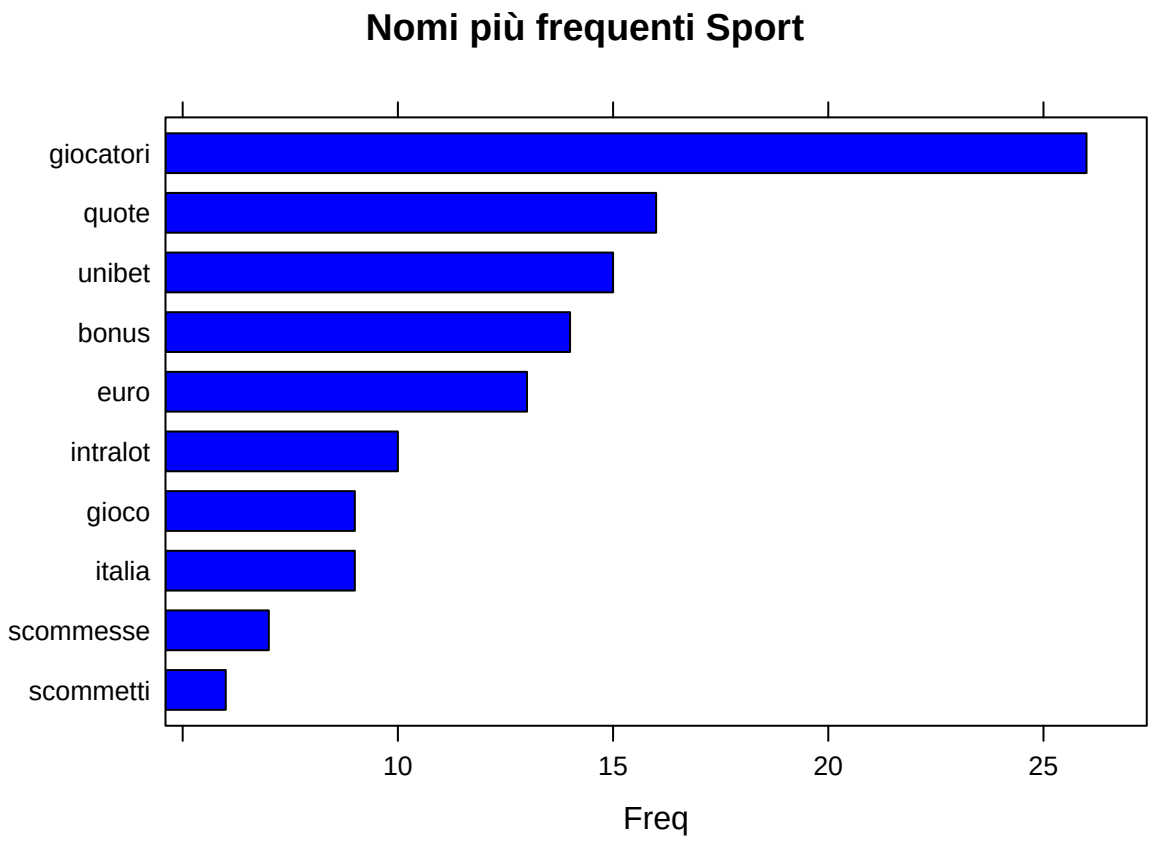


Figure 25: Barplot nomi sport

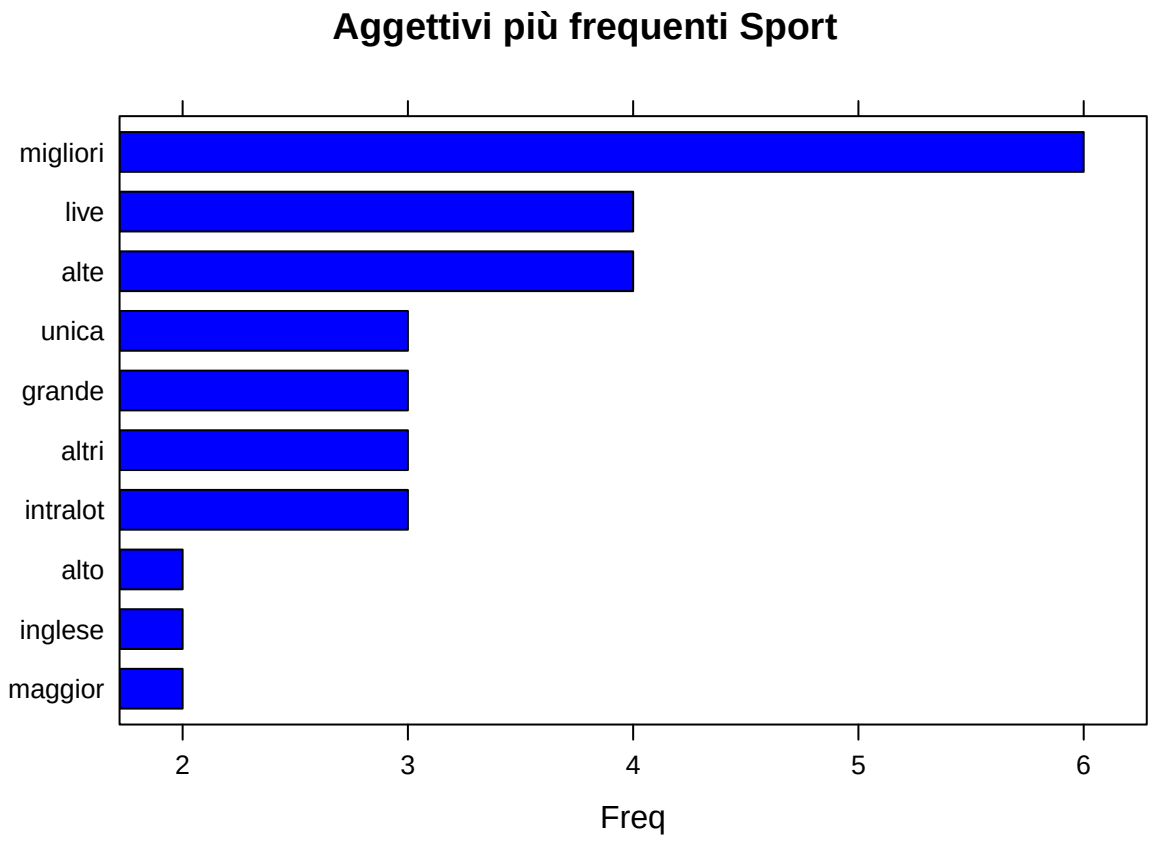


Figure 26: Barplot aggettivi sport

```
stats_sport_adj <- subset(dataframe_sport, upos %in% c("ADJ"))
stats_sport_adj <- txt_freq(stats_sport_adj$token)
stats_sport_adj$key <- factor(stats_sport_adj$key,
                              levels = rev(stats_sport_adj$key))
barchart(key ~ freq, data = head(stats_sport_adj, 10), col = "blue",
         main = "Aggettivi più frequenti Sport",
         xlab = "Freq")
```

```
stats_sport_verb <- subset(dataframe_sport, upos %in% c("VERB"))
stats_sport_verb <- txt_freq(stats_sport_verb$token)
stats_sport_verb$key <- factor(stats_sport_verb$key,
                               levels = rev(stats_sport_verb$key))
barchart(key ~ freq, data = head(stats_sport_verb, 10), col = "blue",
         main = "Verbi più frequenti Sport",
         xlab = "Freq")
```

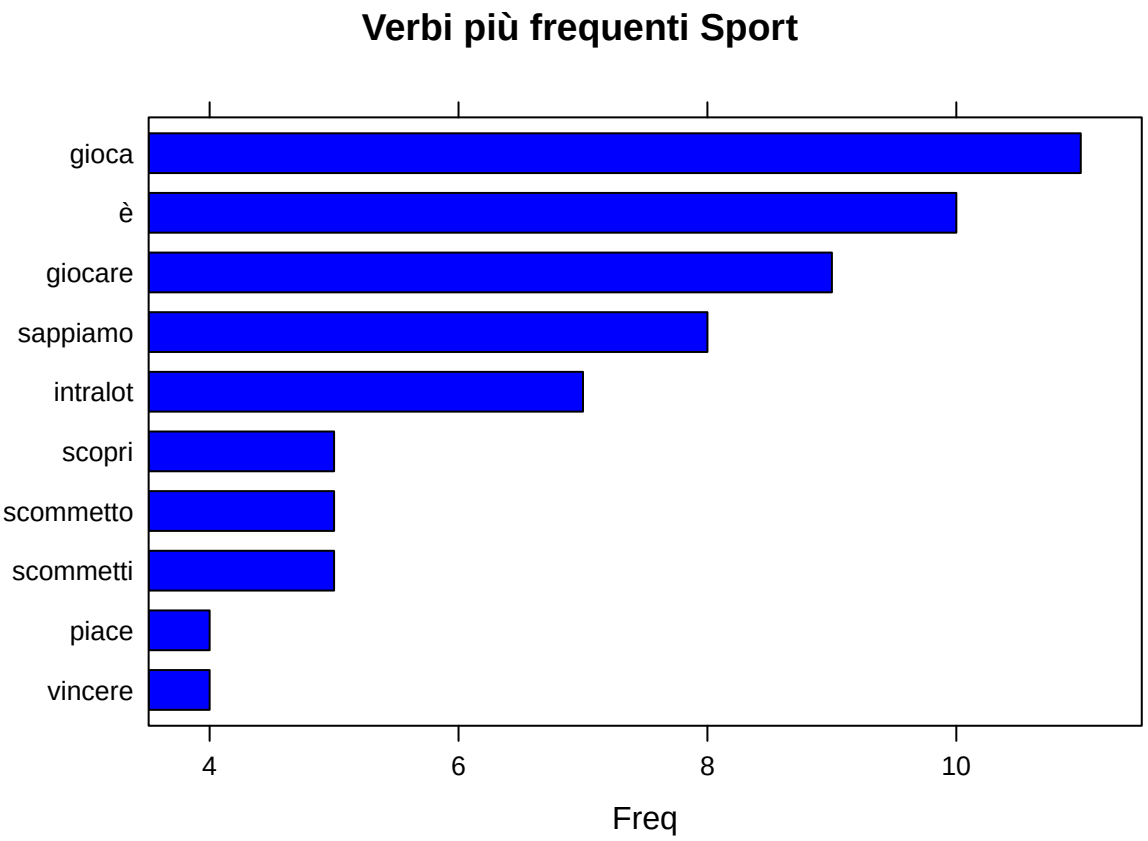


Figure 27: Barplot verbi sport

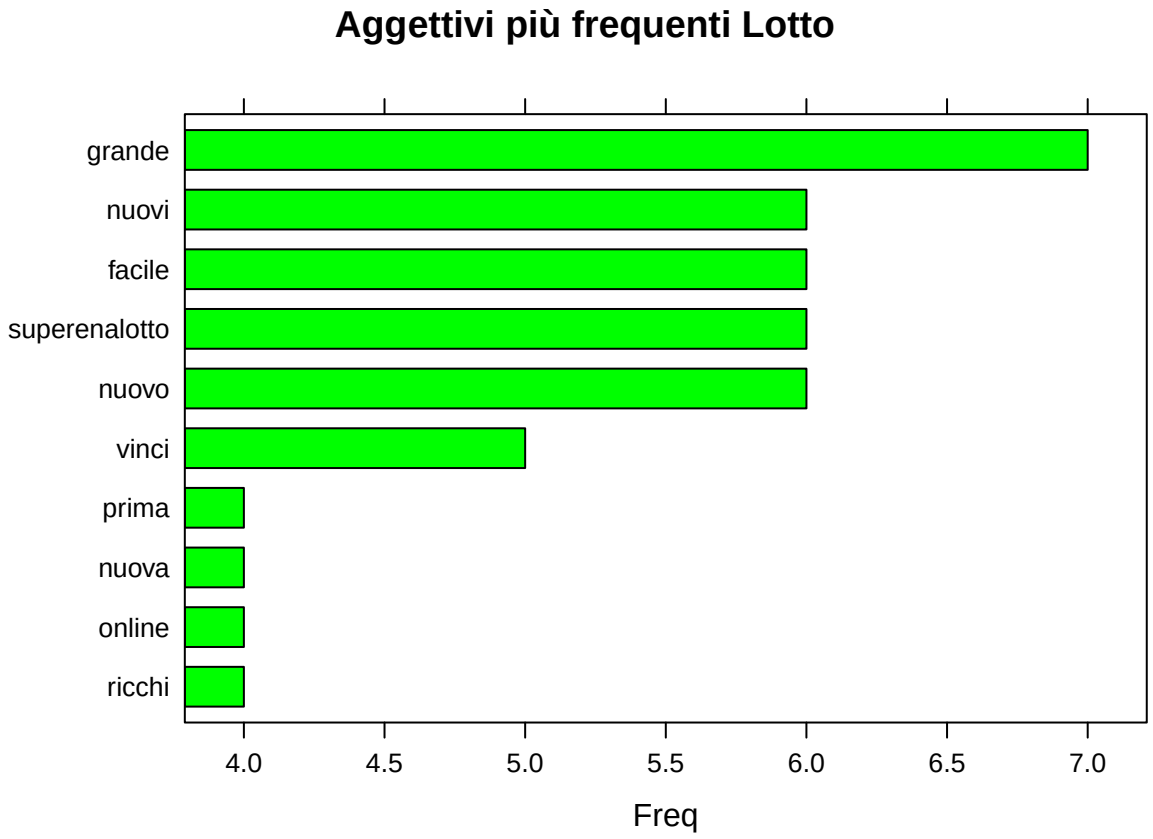


Figure 28: Barplot aggettivi lotto

```
stats_lotto_noun <- subset(dataframe_lotto, upos %in% c("NOUN"))
stats_lotto_noun <- txt_freq(stats_lotto_noun$token)
stats_lotto_noun$key <- factor(stats_lotto_noun$key,
                               levels = rev(stats_lotto_noun$key))
b_l_n<-barchart(key ~ freq, data = head(stats_lotto_noun, 10),col = "green",
                main = "Nomi più frequenti Lotto",
                xlab = "Freq")
```

```
stats_lotto_adj <- subset(dataframe_lotto, upos %in% c("ADJ"))
stats_lotto_adj <- txt_freq(stats_lotto_adj$token)
stats_lotto_adj$key <- factor(stats_lotto_adj$key,
                              levels = rev(stats_lotto_adj$key))
barchart(key ~ freq, data = head(stats_lotto_adj, 10), col = "green",
               main = "Aggettivi più frequenti Lotto",
               xlab = "Freq")
```

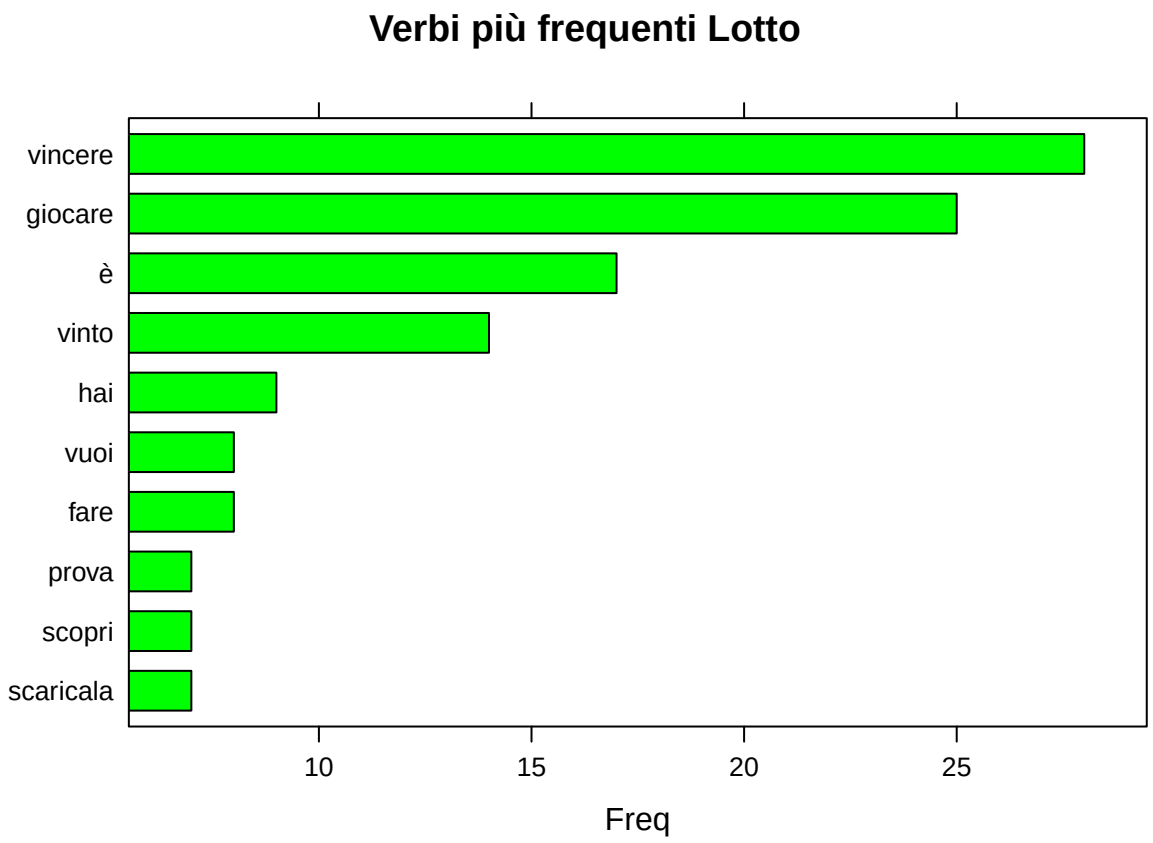


Figure 29: Barplot verbi lotto

```
stats_lotto_verb <- subset(dataframe_lotto, upos %in% c("VERB"))
stats_lotto_verb <- txt_freq(stats_lotto_verb$token)
stats_lotto_verb$key <- factor(stats_lotto_verb$key,
                               levels = rev(stats_lotto_verb$key))
barchart(key ~ freq, data = head(stats_lotto_verb, 10), col = "green",
         main = "Verbi più frequenti Lotto",
         xlab = "Freq")
```

3.4 Discussione

L'intento della ricerca era quello di prendere in esame i testi pubblicitari del gioco d'azzardo ed esplorarli per accertare se ci fossero dei pattern ricorrenti nella comunicazione adottati al fine di persuadere i consumatori a giocare. Ad una rapida lettura dei messaggi trascritti, ancor prima di essere analizzati con tecniche avanzate, emerge immediatamente la brevità e concisione della comunicazione. Questa caratteristica si riflette sui primi

indici esplorativi presi in esame. Il TTR è particolarmente emblematico: i punteggi sono piuttosto bassi per `poker.txt`(0.45) e `lotto.txt`(0.45); unica eccezione `sport.txt`(0.57), ma il punteggio non è particolarmente indicativo di una varietà lessicale così elevata. Dallo studio di frequenza sono stati estratti i temi più rilevanti di ciascuna categoria. Le pubblicità in stile casinò prediligono calcare sul termini 'gioco', 'gioca', 'giochi' e pone l'accento sull'attività online. Le scommesse sportive invece risultano piuttosto tautologiche e mettono in risalto i termini derivanti da 'scommessa'. Lotto, bingo e affini sono più orientati nel porre in risalto la possibilità di vincere. Le reti semantiche sono i grafici che aiutano di più a comprendere la struttura globale di ciascuna categoria di pubblicità . Per il documento `poker` il fulcro dei temi è dominato dalla triade 'gioco-casinò-euro', interconnesse ad una serie di altre tematiche come la vittoria e il gioco online. Il documento `sport` vede come chiave centrale il termine 'giocatori', connesso con una serie di elementi relativi alle scommesse e alle vincite in denaro. `lotto` invece vede una rete più intricata che ruota attorno al concetto dei numeri che permettono che permettono di vincere, con l'apporto di una buona dose di fortuna. Lo studio di similarità tra `sport` e `lotto` non è stato molto fruttuoso: le tematiche in comune sono relative al rapporto gioco-giocatori e vincite (nulla di sorprendente). Utilizzando la LDA sulle parole non in comune del documento `sport` i topic emersi riflettono quanto già desunto dal semplice studio di frequenza: il ruolo cruciale dell'azione di scommettere. La LDA sul documento `lotto` invece fa emergere dei topic più variegati: oltre al tema focale in cui si incentrano questa tipologia di giochi (l'utilizzo dei numeri), emergono una serie di termini che potrebbero essere ricondotti alla vita quotidiana o alle aspirazioni di vita ('sogno', 'sogna', 'amore', 'mamma', 'papà'). Tramite l'UPOS è possibile riscontrare quali siano le categorie lessicali più utilizzate in ciascun contesto. Sono tre le categorie su cui porre maggior attenzione: nomi, aggettivi, verbi. Per tutti e tre i documenti i nomi risultano essere in assoluto la categoria più inflazionata. I verbi e gli aggettivi sono utilizzati in maniera proporzionalmente diversa, ma per tutte e tre i documenti i verbi hanno una frequenza maggiore di utilizzo rispetto agli aggettivi. Forse gli unici dettagli degni di nota che emrgono sono il fatto che nella categoria aggettivi il documento `poker` presenta un'alta frequenza di richiami al mondo digitale rispetto agli altri documenti, mentre nella categoria verbi il testo `lotto` predilige l'uso della forma verbale 'vincere'. In conclusione l'analisi statistica testuale applicata alle pubblicità del gioco non ha dato i risultati sperati, facendo emergere pochi dettagli rispetto a quanto ci si aspettasse. I limiti della ricerca possono essere ricondotti alla lunghezza dei documenti non molto elevata, dovuta probabilmente ad una scelta dei brand pubblicitari di redarre pubblicità semplici e concise; bisogna tener conto inoltre che la legge pone dei limiti piuttosto vincolanti su cosa può e non può essere detto (Decreto Balduzzi). La *Sentiment Analysis* sarebbe stata un lavoro interessante da applicare. Purtroppo i tentativi che sono stati svolti non sono andati a buon fine, in quanto i dizionari italiani utilizzati non sono sufficientemente implementati per poter eseguire un'operazione di questo genere con sufficiente accuratezza. Per futuri lavori di applicazione di questa tecnica sul gioco d'azzardo suggerisco o l'ampliamento dei dati di base utilizzando anche gli annunci publicitari presenti nel web, oppure spostare il focus d'attenzione dai brand del gioco ai commenti dei giocatori presenti nei social network e blog specializzati.

Bibliografia

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: {DSM-5}*. 5th ed. Washington, {DC}: Autor.

Anderson, G., and R. I.F. Brown. 1984. "Real and laboratory gambling, sensation-seeking and arousal." PhD thesis. doi:10.1111/j.2044-8295.1984.tb01910.x.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://cran.r-project.org/package=gridExtra.

Baker, Frank B. 1974. "Stability of Two Hierarchical Grouping Techniques Case I: Sensitivity to Data Errors." *Journal of the American Statistical Association* 69 (346). Taylor & Francis: 440–45. doi:10.1080/01621459.1974.10482971.

Belkin, Nicholas J., and W. Bruce Croft. 1992. "Information filtering and information retrieval: two sides of the same coin?" *Communications of the ACM* 35 (12): 29–38. doi:10.1145/138859.138861.

Benoit, Kenneth, and Adam Obeng. 2019. *readtext: Import and Handling for Plain and Formatted Text Files*. https://cran.r-project.org/package=readtext.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software* 3 (30): 774. doi:10.21105/joss.00774.

Bergler, E. 1957. *The psychology of gambling*. Edited by Hill & Wang.

Blei, David M, Blei@cs Berkeley Edu, Andrew Y Ng, Ang@cs Stanford Edu, Michael I Jordan, and Jordan@cs Berkeley Edu. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Camilo, Co, Jc Silva, Daniele El-jaick, Tayena Hendrickx, Boris Cule, Pieter Meysman, Stefan Naulaerts, et al. 2015. "Mining association rules in graphs based on frequent cohesive itemsets." PhD thesis. doi:10.1007/978-3-319-18032-8_50.

Clark, Luke. 2010. "Decision-making during gambling: An integration of cognitive and psychobiological approaches." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1538): 319–30. doi:10.1098/rstb.2009.0147.

Feinerer, Ingo, Kurt Hornik, and David Meyer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25 (5): 1–54. http://www.jstatsoft.org/v25/i05/.

Freud, S. 1928. *Dostoyevsky and parracide. In J. Strachey (Ed.)*. Edited by Hogarth Press. Vol. The standa.

Galili, Tal. 2015. "dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical

clustering." *Bioinformatics*. doi:10.1093/bioinformatics/btv428.

Griffiths, Mark D. 1990. "The cognitive psychology of gambling." *Journal of Gambling Studies* 6 (1): 31–42. doi:10.1007/BF01015747.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with {lubridate}." *Journal of Statistical Software* 40 (3): 1–25. http://www.jstatsoft.org/v40/i03/.

Grün, Bettina, and Kurt Hornik. 2011. "{topicmodels}: An {R} Package for Fitting Topic Models." *Journal of Statistical Software* 40 (13): 1–30. doi:10.18637/jss.v040.i13.

Jabr, F. 2013. "Gambling on the Brain." PhD thesis, Nature Publishing Group. doi:10.1038/scientificamerican1113-28.

Jockers, Matthew L. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*. https://github.com/mjockers/syuzhet.

Kahneman, Daniel, and Amos Tversky. 1982. "The psychology of preferences." *Scientific American* 246 (1). US: Scientific American, Inc.: 160–73. doi:10.1038/scientificamerican0182-160.

King, Brandy E., and Kathy Reinold. 2014. "Natural language processing." PhD thesis. doi:10.1016/b978-1-84334-318-9.50005-3.

Langer, Ellen J. 1975. "The illusion of control." *Journal of Personality and Social Psychology* 32 (2). US: American Psychological Association: 311–28. doi:10.1037/0022-3514.32.2.311.

Liu, Bing. 2012. "Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers." PhD thesis. doi:10.1007/978-1-4899-7502-7_907-1.

Loftus, Geoffrey R, and Elizabeth F Loftus. 1983. *Mind at Play; The Psychology of Video Games*. New York, NY, USA: Basic Books, Inc.

Lopez-Gonzalez, Hibai, Frederic Guerrero-Solé, and Mark D Griffiths. 2018. "A content analysis of how 'normal' sports betting behaviour is represented in gambling advertising." *Addiction Research and Theory* 26 (3). Informa UK Ltd.: 238–47. doi:10.1080/16066359.2017.1353082.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2019. *cluster: Cluster Analysis Basics and Extensions*.

May, Ryan K, James P Whelan, Timothy A Steenbergh, and Andrew W Meyers. 2003. "The gambling self-efficacy questionnaire: an initial psychometric evaluation." PhD thesis. http://www.ncbi.nlm.nih.gov/pubmed/14634297.

Moscovici, S. 2015. *La psychanalyse, son image et son public*. Bibliothèque de Psychanalyse. Presses

Universitaires de France. https://books.google.it/books?id=RFVfCwAAQBAJ.

Moscovici, Serge. 1984. “The Phenomenon of Social Representations.” In *Social Representations*, 2:3–69.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org.

Reid, R. L. 1986. “The psychology of the near miss.” *Journal of Gambling Behavior* 2 (1): 32–39. doi:10.1007/BF01019932.

S, Vijayarani, and Janani R. 2016. “Text Mining: open Source Tokenization Tools – An Analysis.” PhD thesis. doi:10.5121/acii.2016.3104.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer. http://lmdvr.r-forge.r-project.org.

Sharpe, Louise. 2002. “A reformulated cognitive - Behavioral model of problem gambling: A biopsychosocial perspective.” *Clinical Psychology Review* 22 (1): 1–25. doi:10.1016/S0272-7358(00)00087-8.

Silge, Julia. 2017. *janeaustenr: Jane Austen’s Complete Novels*. https://cran.r-project.org/package=janeaustenr.

Silge, Julia, and David Robinson. 2016. “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *JOSS* 1 (3). The Open Journal. doi:10.21105/joss.00037.

Skinner, B.F. 1953. “Science and human behavior. Macmillan: New York.” PhD thesis.

Spina, Stefania, and Scienze Umane. 2014. “Il Perugia Corpus : una risorsa di riferimento per l ’ italiano . Composizione , annotazione e valutazione,” 354–59.

Tan, Ah-Hwee. 1999. “Text mining: The state of the art and the challenges.” *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* 8: 65–70.

Vijayarani, S., J. Ilamathi, and Nithya. 2018. “Preprocessing Techniques for Text Mining - An Overview.” PhD thesis. doi:10.1016/j.procs.2013.05.286.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2019. *dplyr: A Grammar of Data Manipulation*. https://cran.r-project.org/package=dplyr.

Wijffels, Jan. 2019. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the ’UDPipe’ ’NLP’ Toolkit*. https://cran.r-project.org/package=udpipe.

Xie, Yihui. 2014. “knitr: A Comprehensive Tool for Reproducible Research in {R}.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D Peng.

Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhao, Shilin, Linlin Yin, Yan Guo, Quanhu Sheng, and Yu Shyr. 2019. *heatmap3: An Improved Heatmap Package*. https://cran.r-project.org/package=heatmap3.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort.* Oxford, England: Addison-Wesley Press.