DMIF, Università di Udine

Tecnologie Digitali per il Cibo e la Ristorazione

Text Mining

Andrea Brunello
andrea.brunello@uniud.it

A.A. 2021–2022

Outline

1. Introduzione
2. Pre-processamento del testo
3. Applicazioni del Text Mining
4. Written Notes Evaluation at Gap Srlu (optional)

Introduzione

Definizione di text mining

Il text mining può essere definito come:

Il processo di estrarre nuova informazione utile, precedentemente sconosciuta, da diverse fonti di testo.

Le fonti testuali possono essere, ad esempio, siti web, libri, e-mail, recensioni, articoli.

Il text mining implica, quando vengono utilizzati per esso approcci di machine learning di tipo classico: **il processo di strutturazione del testo in input**, la derivazione di modelli a partire dai dati strutturati e, infine, la valutazione e l'interpretazione dell'output.

Fondamenti sul testo

Per i nostri fini, definiamo il *testo* come un tipo di dato non strutturato costituito da una serie di paragrafi, ciascuno composto da una o più frasi, ognuna formata da stringhe chiamate parole.

Le *frasi* iniziano con una lettera maiuscola e terminano con un punto. Una frase può avere una o più *clausole*, che possono essere unite tra loro da congiunzioni o da pronomi relativi.

Un *paragrafo* è un elenco ordinato di frasi coerenti con un particolare sottoargomento. I paragrafi iniziano con un rientro e terminano con un ritorno a capo.

La *parola* è considerata come l'unità base di testo. Tuttavia, a volte le parole "vanno insieme", ad esempio: possono portare, oltre che, al contrario, spreco alimentare, problema serio, ...

Formati del testo

Il testo può essere archiviato ed elaborato utilizzando formati diversi, il che aumenta la difficoltà delle attività di estrazione delle informazioni:

- MS Word
- MS Powerpoint
- PDF Adobe
- XML
- HTML
- Plain text

Pre-processamento del testo

Pre-processamento del testo

Poiché il testo è un tipo di dato non strutturato, è difficile analizzarlo nella sua forma originale, a meno di non ricorrere a reti neurali.

Per applicare gli approcci di analisi classici, dobbiamo descriverlo mediante un numero fisso di attributi.

In questo senso, il primo passo è quello di svolgere il *text indexing*, cioè convertire il testo in una lista di parole. In seguito, dobbiamo determinare una strategia per assegnare un punteggio di importanza a ciascuna parola, un compito chiamato *term weighting*.

Obiettivo finale

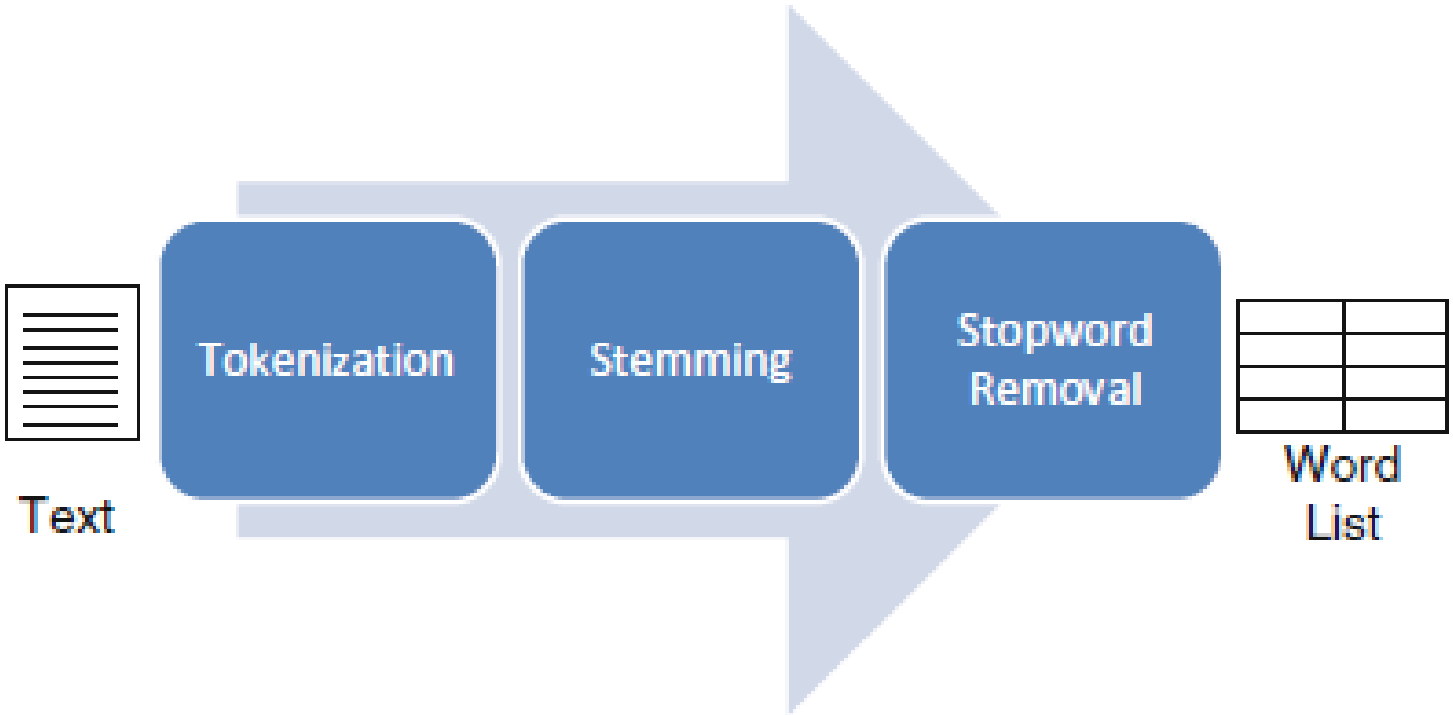
Il fine ultimo è definire un insieme fisso di attributi.

Ogni testo (documento) presente nel dataset (corpus o raccolta) viene descritto da tale insieme di attributi.

Come vedremo, gli attributi consentono di esprimere informazioni relative al contenuto dei testi.

Su tale dataset tabellare è poi semplice applicare le tecniche di data mining viste nelle scorse lezioni.

I tre passi del text indexing



Tokenizzazione

Il processo di tokenizzazione prevede la segmentazione del testo in token.

I token vengono identificati mediante gli spazi bianchi, la punteggiatura e i caratteri speciali.

Le parole con uno o più caratteri speciali possono essere eliminate del tutto, o possono essere rimossi solo i caratteri speciali (ad es. 25%).

Le parole vengono trasformate in minuscolo, in modo da ottenere una rappresentazione univoca per ogni parola.

Esempio di tokenizzazione

Text categorization refers to the process of assign a category or some categories among predefined ones to each document, automatically. Text categorization is a pattern classification task for text mining and necessary for efficient management of textual information systems.

Tokens
text
categorization
refers
to
the
process
of
assign
a
category
.....

Stemming

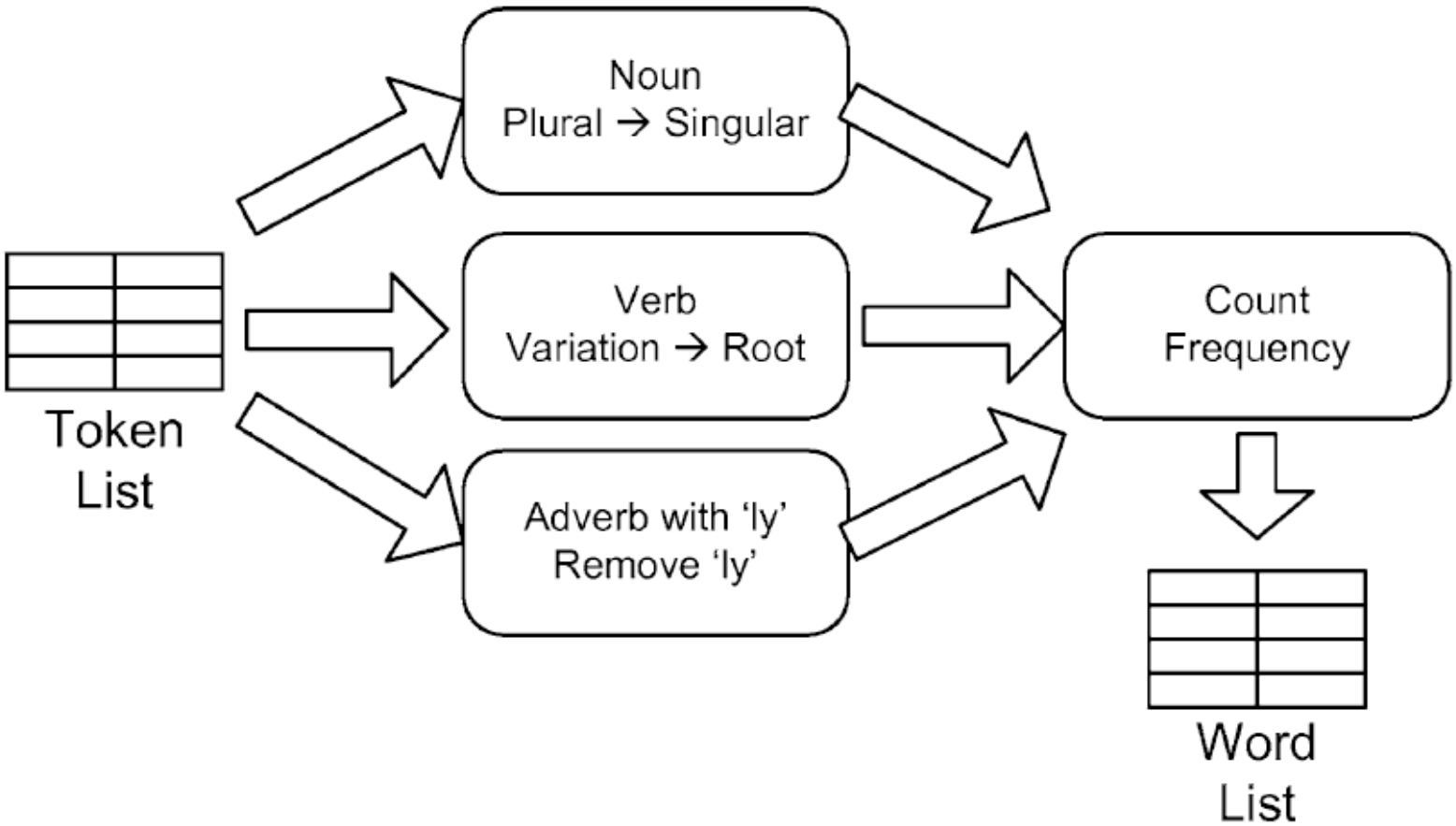
Lo stemming è il processo di conversione di ogni token ottenuto dal passaggio precedente nella sua forma radice.

Lo stemming è generalmente applicabile a nomi, verbi e aggettivi.

Il risultato di questo processo è un elenco di parole nella loro forma radice.

Vengono applicate delle regole specifiche della lingua presa in considerazione (ad es. Algoritmo di Porter o di Snowball).

Processo di stemming



Esempi di stemming

Varied Form	Root Form
better	good
best	good
simpler	simple
simplest	simple
assigning	assign
assigned	assign
assignment	assign
complexity	complex
analysis	analyze
categorization	categorize
categorizing	categorize
categorizes	categorize

Rimozione delle stopword

La rimozione delle stopword è il processo di rimozione delle parole non significative dall'elenco dei token.

Le stopword sono definite come le parole più comuni in una lingua, anche se non esiste un consenso universale su di esse.

Alcuni esempi sono le preposizioni, come 'in', 'su', 'per' e così via. Congiunzioni come 'e', 'o', 'ma', 'comunque'.

Intuitivamente, si desidera rimuovere le stopword poiché non portano alcuna informazione, essendo troppo comuni.

Caveat: in alcuni casi, le stopword possono effettivamente essere importanti. Prendiamo ad esempio il gruppo rock inglese *The Who*. Oppure, la frase *Ti ho detto che non era felice*, che potrebbe risultare in ['detto', 'felice '].

Dalle parole agli n-grammi

Trattare singole parole ha un grosso svantaggio: non è possibile tenere conto del loro ordine e contesto.

Ad esempio, si osservino le seguenti due frasi, che possono sembrare identiche se si prende in considerazione solo l'insieme delle singole parole:

- *It seems you were right not inviting him*
- *It seems you were not right inviting him*

Un *n-gramma* è una sequenza contigua di *n* elementi di un dato testo.

Attraverso gli n-grammi è possibile tenere meglio traccia dei contesti delle parole, o considerare *phrasal verb* come 'get around', 'pass out', 'carry on', 'andare su', 'far fuori'.

Esempi di n-grammi

Full sentence	It does not, however, control whether an exaction is within Congress's power to tax.
Unigrams	"It"; "does"; "not,"; "however,"; "control"; "whether"; "an"; "exaction"; "is"; "within"; "Congress's"; "power"; "to"; "tax."
Bigrams	"It does"; "does not,"; "not, however,"; "however, control"; "control whether"; "whether an"; "an exaction"; "exaction is"; "is within"; "within Congress's"; "Congress's power"; "power to"; "to tax."
Trigrams	"It does not"; "does not, however"; "not, however, control"; "however, control whether"; "control whether an"; "whether an exaction"; "an exaction is"; "exaction is within"; "is within Congress's"; "within Congress's power"; "Congress's power to"; "power to tax."

Term weighting

Il term weighting si riferisce al processo di calcolo e assegnazione di un peso ad ogni parola (o, in generale, n-gramma), in modo da riflettere il suo grado d'importanza nel testo.

Tale importanza è tipicamente legata alla frequenza del termine, che può essere espressa in due modi diversi:

- *frequenza assoluta*: il numero di occorrenze della parola in un testo
- *frequenza relativa*: il rapporto fra il numero di occorrenze della parola e il numero di parole nel testo

Attenzione all'uso delle frequenze

Fare affidamento solo sulle frequenze assolute/relative delle parole non è sufficiente.

Dato un insieme di documenti provenienti dallo stesso dominio, alcune parole sono naturalmente più frequenti di altre (ad esempio, *interesse* in un dominio finanziario).

Data la loro frequenza elevata, tali termini potrebbero non essere molto importanti, perché sono comuni (possono essere intesi come delle stopword legate al dominio).

La strategia TF-IDF cerca di risolvere tale problematica nella determinazione dei pesi.

TF-IDF

Term frequency–inverse document frequency (TF-IDF) è una metrica che ha lo scopo di determinare quanto sia importante una parola per un *documento* in una *raccolta*.

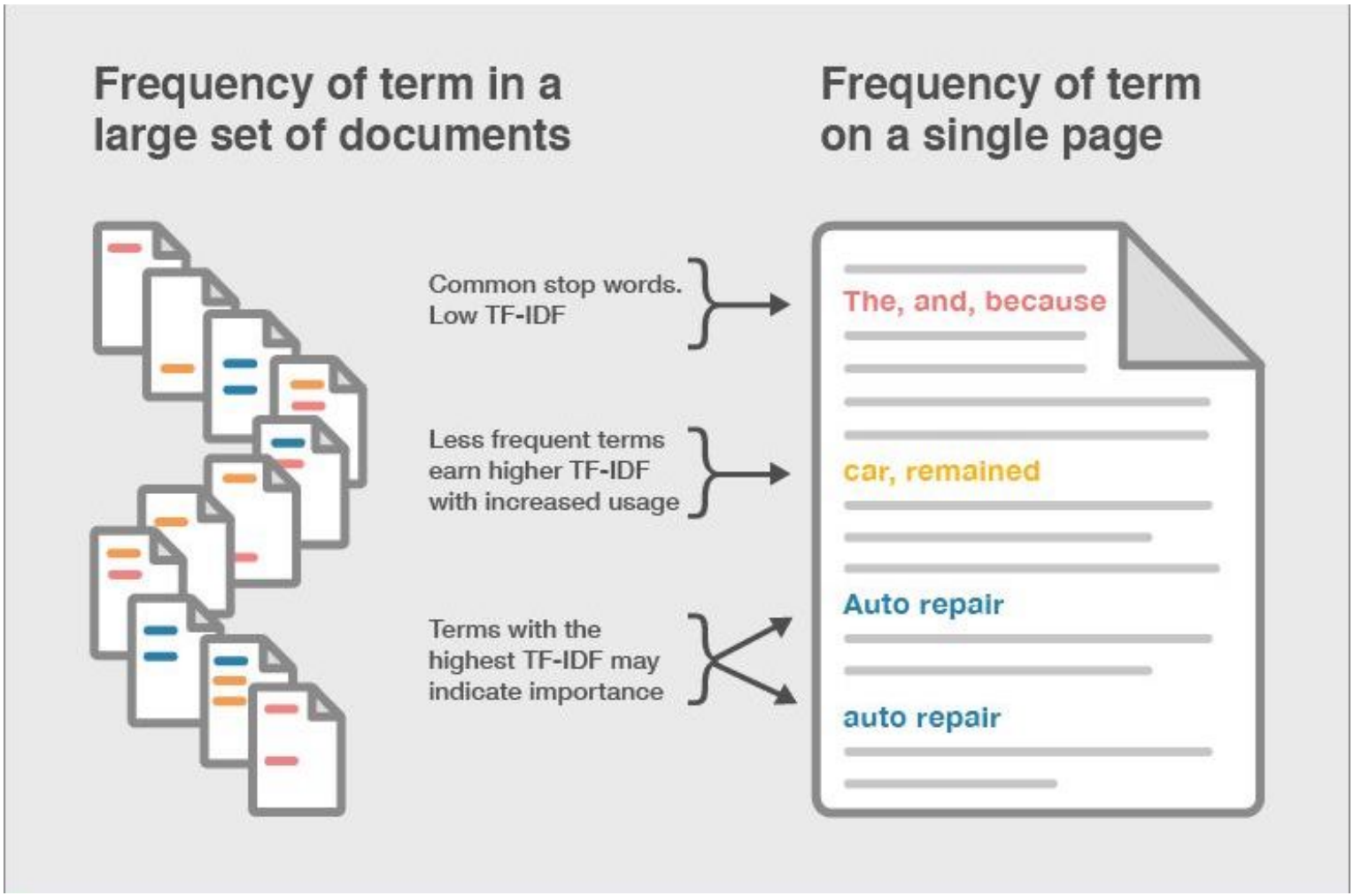
Il suo valore aumenta proporzionalmente al numero di volte che la parola appare nel documento e diminuisce al crescere del numero di documenti nella raccolta che contengono tale parola.

Data una parola t nel documento d, il suo peso W_{td} è dato da:

$$W_{td} = log(\frac{N}{DF_t})TF_{td}$$

dove N è il numero totale di documenti nella raccolta, DF_t è il numero di documenti che includono la parola t e TF_{td} è il numero di occorrenze della parola t nel documento d, diviso la lunghezza di d.

TF-IDF – Esempio



Applicazioni del Text Mining

Applicazioni del Text Mining

Applicazioni tipiche includono:

- Visualizzazione
- Classificazione
- Clustering
- Tagging/annotazione
- Concept/entity extraction
- Sentiment analysis
- Document summarization

In genere, prima di qualsiasi tipologia di analisi, è necessario eseguire comunque le operazioni di pre-processing sul testo.

Visualizzazione

La visualizzazione del testo è un'attività essenziale per eseguire l'*analisi esplorativa dei dati*, vale a dire prendere confidenza con i dati prima di addentrarsi in attività di analisi più complesse.

Inoltre, visualizzando il testo è possibile, a colpo d'occhio:

- comprendere il contenuto generale del documento
- raggruppare documenti
- confrontare documenti

Alcune tipologie di visualizzazione sono intuitive (ad esempio, un istogramma che mostra la frequenza delle parole), altre sono più complesse (word cloud, alberi, . . .).

Testo d'esempio

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We wer losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of yo

Word cloud

Le word cloud mostrano, in modo molto intuitivo, i termini che appaiono più frequentemente in un documento.

Word cloud – 2


people
will
care
can
health
system
insurance
Americans
doctors
American
know
want
year
medical
money
cost
small
every
right
business
government
coverage
plan
children
still
must
also
better
sick
say
pay
make
savings
time
just
change
believe
quality
businesses
tonight
new
think
country
responsibility
provide
get
first
together
large
years
Now
Bill Clinton 1993

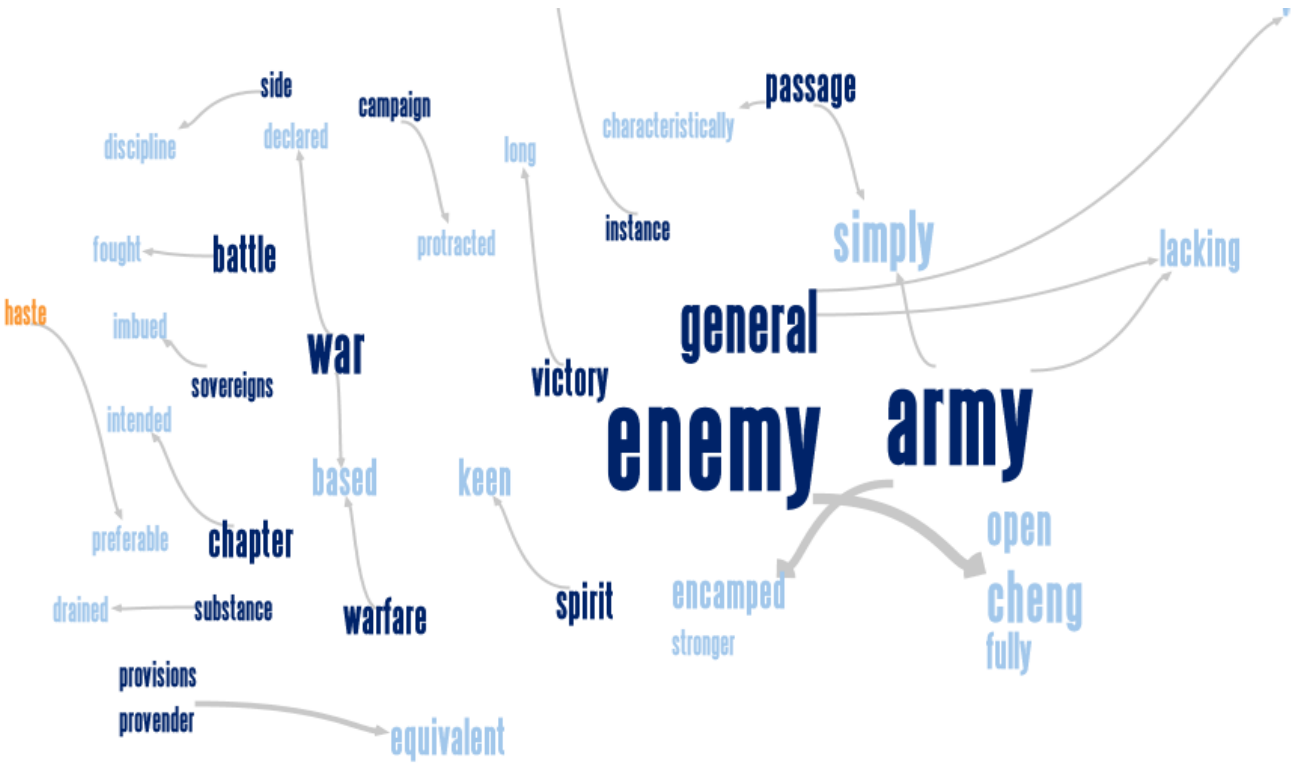

insurance
health
will
care
coverage
companies
plan
government
Americans
Medicare
businesses
many
still
believe
know
time
people
costs
provide
one
reform
system
affordable
public
every
pay
better
without
small
country
option
us
make
also
American
need
part
work
security
can
get
year
like
help
offer
Barack Obama 2009

Word tree

Raggruppano le frasi che iniziano con una stessa sequenza di parole.

Phrase net

Può anche essere interessante considerare coppie di parole che sono collegate fra loro tramite qualche altra parola. Ad esempio, queste sono alcune delle parole collegate dal termine *is* in *L'arte della guerra (Sun Tzu)*.



Classificazione / Regressione

Classificazione è il processo di assegnare a ciascuna istanza una o più categorie tra quelle predefinite (valori discreti).

Nei task di *regressione*, invece, viene assegnato un numero (il dominio è continuo).

Considerando il testo, si potrebbe voler assegnare una categoria a ciascun documento, ad esempio l'elenco degli argomenti trattati in esso; oppure, potremmo voler determinare il punteggio complessivo del sentiment del testo.

Esercizio: IMDB sentiment dataset

- Consideriamo il dataset *imdb_sentiment.arff*, contenente il testo di 2000 recensioni di film, equamente divise fra le due classi *positiva* e *negativa*
- Vogliamo costruire un modello interpretabile in grado di prevedere il sentiment di una recensione
- Il primo step è effettuare il pre-processamento del testo in modo da portarlo in una forma strutturata (tokenizzazione, stemming, calcolo dei pesi)
- A tal fine, selezioniamo il filtro *unsupervised* > *attribute* > *StringToWordVector*, applicandolo al primo attributo
- In esso, impostiamo *IDFTransform* = *True*, *minTermFreq* = *100*, *lowerCaseTokens* = *True*, *stemmer* = *Snowball*, *stopwordsHandler* = *Rainbow*
- Infine, costruiamo un albero di decisione sul dataset, impostando *Percentage split* = *90%* su *Test options*

Written Notes Evaluation at Gap Srlu

Written notes evaluation at Gap Srlu

Tracking the performance of agents is a primary issue in contact centers, as it allows, for example:

- the best match to be taken between service and agent
- the recognition of unsatisfactory agent behaviours, due for example to a lack of proper training
- the prediction of future trends, based on the history of observations

As a part of its agent performance evaluation framework, Gap automatically assesses the quality of written notes taken by the agents during phone calls.

Written notes analysis

Idea:

- how often / in which way does an agent record notes regarding an inbound call?
- compare single agent behaviour with service average values

How to evaluate written notes?

- extract summarizing features from the text
- identify groups of similar notes
- devise a methodology to assign a generic new note to one of the groups

Feature extraction from plain text

For each note, we calculate (*R* script, *openNLP* library):

- numbers of words and characters
- *Gulpease* readability index value
- fractions of articles and conjunctions over words
- fractions of verbs and adverbs over words
- fraction of adjectives over words
- fraction of prepositions over words
- fraction of quantifiers over words
- fraction of (pro)nouns over words
- fraction of numeric codes over words
- fraction of proper nouns over words
- fraction of words/abbreviations found in Italian dictionary
- fraction of words found in **service-specific** domain
- fraction of unrecognized words

Identify groups of similar notes

- Random sampling of 1000 notes
- application of a clustering algorithm to the selected notes (*E-M* algorithm)
- 6 clusters emerged:
  - articulated notes
  - non-articulated notes
  - abbreviated notes
  - domain-specific notes
  - nonsense notes
  - hybrid notes

Classify a new note

- Attach a new feature to each of the clustered notes: *cluster label*
- apply a decision tree learning algorithm (*J48*), with the goal of predicting the label (94.7% accuracy)
- the tree can then be used to classify new notes

```
riconosciuti_abbr_su_parole <= 0.142857
|   riconosciuti_dominio_su_parole <= 0.133333
|   |   preposizioni_su_parole <= 0
|   |   |   non_riconosciuti_su_parole <= 0.157895
|   |   |   |   congiunzioni_su_parole <= 0.025
|   |   |   |   |   articoli_su_parole <= 0.071429: non_articulated_notes
|   |   |   |   |   articoli_su_parole > 0.071429: articulated_notes
|   |   |   |   congiunzioni_su_parole > 0.025: articulated_notes
|   |   |   non_riconosciuti_su_parole > 0.157895
|   |   |   |   non_riconosciuti_su_parole <= 0.333333
|   |   |   |   |   articoli_su_parole <= 0.083333: hybrid_notes
|   |   |   |   |   articoli_su_parole > 0.083333: articulated_notes
|   |   |   |   non_riconosciuti_su_parole > 0.333333: non_sense_notes
|   |   preposizioni_su_parole > 0
|   |   |   indice_gulp <= 129.833333: articulated_notes
|   |   |   indice_gulp > 129.833333
|   |   |   |   non_riconosciuti_su_parole <= 0.0625: non_articulated_notes
|   |   |   |   non_riconosciuti_su_parole > 0.0625: hybrid_notes
```

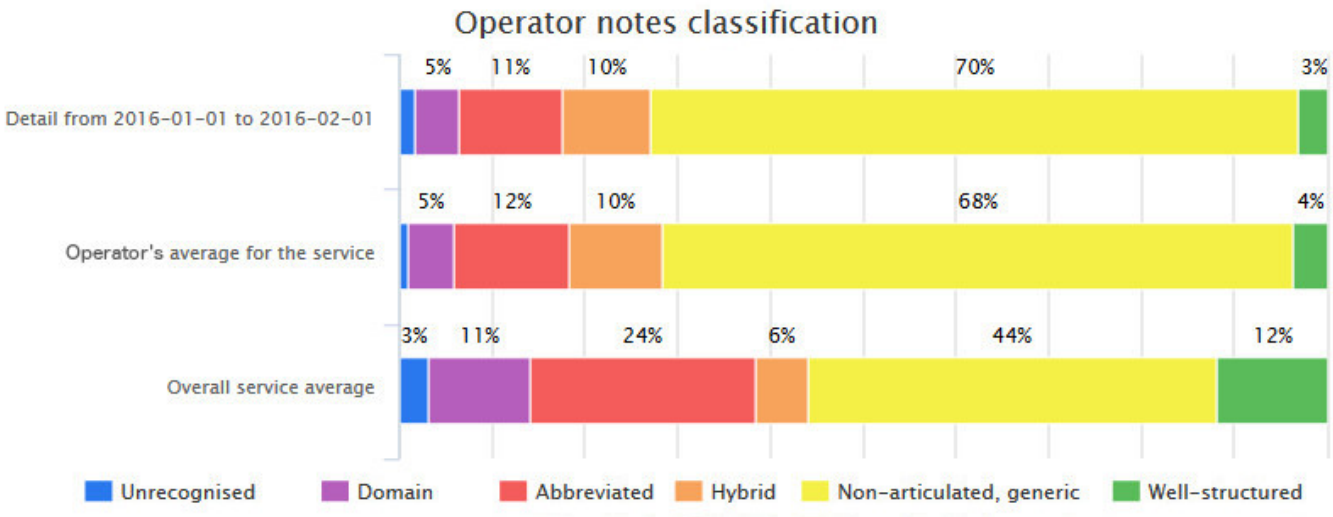
Example

	valore text	gruppo_nota text
1	info voltura	hybrid
2	invio del f24 unico	articulated
3	informazioni per appunt sub e comunica dati catastali	articulated
4	info posizione pagamenti mensa scolastica	hybrid
5	NON RISPONDE	non-articulated
6	Info	abbreviated
7	VIA [redacted] MQ 37 C'è SCRITTO 43 BOLLETTAZIONE SBAGLIATA. DEVE PASSARE AGLI SPORTELLI PER RETTIFICA DI METRATURA CON PIANTINA SCALA 1:100. RIFERISCO. C	articulated
8	SIGNORA CHIAMA PER SAPERE SE è STATA APPLICATA LA DETRAZIONE DI 25 euro per figlio sul calcol	articulated
9	la signora avea chiamato il 23/05 per una verifica posizione per la TARES: ha un locale comme	articulated
10	chiede quanto deve pagare per la tassa. Parlato con esatto: deve pagare 61 euro.	articulated
11	info boll	abbreviated
12	rimborso ud	non-articulated
13	tasi	domain-specific
14	INFO GENERICHE IMU, TASI	domain-specific
15	info su avv sosp	hybrid
16	chiede se può rateizzare l'importo da versare per la mensa. Riferito che deve fare richiesta	articulated
17	invio copia boll	hybrid
18	chiede il saldo mensa. Riferito che abbiamo problemi tecnici tecnici al server	articulated

- Determine the distribution of note classes for each service
- Then, compare with agent-service distributions

Example – 2

Agent-service notes class distribution, with respect to the overall distribution for the service.



Bibliography

Taeho Jo Text Mining Concepts, Implementation, and Big Data Challenge, Volume 45, Springer