

ITEC874 — Big Data Technologies

Week 8 Lecture 1: Analysing Unstructured Data

Diego Mollá

Department of Computer Science
Macquarie University

COMP348 2018H1

Programme

- 1 Analysing Unstructured Data
- 2 Analysing Text Data

Reading

- Lecture notes.
- Text Analytics — Microsoft Azure Machine Learning Studio.
- Text Analytics Using SAS Text Miner.

Tentative Outline of ITEC874 — Weeks 7 to 12

- 7 Analysing Big Data
- 8 Analysing Unstructured Data
- 9 Visualising Big Data
- 10 Analysing Streaming Data
- 11 Big Data and Society
- 12 Industry Talk: Amazon

Programme

1 Analysing Unstructured Data

2 Analysing Text Data

Why Analyse Unstructured Data

It's about Variety

- Probably the biggest impact of Big Data in companies is the possibility to analyse unstructured data.
- Unstructured data contains information that can potentially be very useful.
- It opens up the possibility to access yet untapped information from multiple sources.

Sources of Unstructured Data

Video: surveillance cameras, videos in social media.

Images: Web images, images in social media, satellite images.

Sound: Call centre recordings.

Text: Documents, reports, webpages, social media posts.

Motivating Example: United Healthcare

- Have recorded voice files from customer calls to call centres.
- The voice data was converted to text using speech-to-text conversion tools.
- The text was then analysed using natural language processing software.
- Their analysis focused on identifying customers who use terms suggesting strong dissatisfaction.

Use Cases of Image Analytics

<https://www.zencos.com/blog/5-applications-of-image-analytics-with-SAS-Viya/>

- 1 Identify bags at airports.
- 2 Analyse social media images for missing persons.
- 3 Real-time vehicle damage assessment.
- 4 Detect pneumonia from chest x-rays.

Programme

1 Analysing Unstructured Data

2 Analysing Text Data

- Characteristics of Text
- Common Building Blocks for Text Analytics
- Some APIs for Text Analytics

Why Analysing Text?

Information Overload

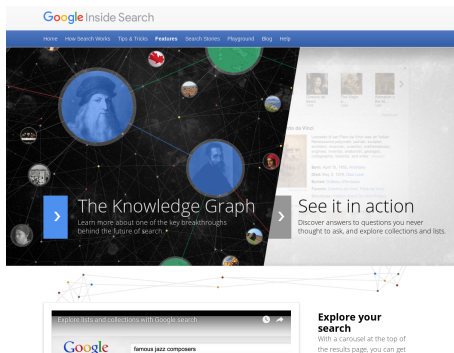
- A lot of information is available as free text.
- The most natural form to write information is through free text.
- A great deal of digital information is available as free text.
- People can read and understand free text easily.
- But it's very hard for machines!



Integrating Natural Language Processing and Data Mining


Results to queries asked in current search engines may be enriched with information mined from:

- Knowledge sources such as Google's Knowledge Graph.
- Text mining based on the characteristics of the query.



<https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html>

Google Search (13 Feb 2018)

 GLE

language technology

[All](#) [News](#) [Images](#) [Videos](#) [Maps](#) [More](#) [Settings](#) [Tools](#)

About 553,000,000 results (0.66 seconds)

Language technology - Wikipedia
https://en.wikipedia.org/wiki/Language_technology ▼
Language technology, often called human language technology (HLT), consists of natural language processing (NLP) and computational linguistics (CL) on the one hand, and speech technology on the other. It also includes many application oriented aspects of these.

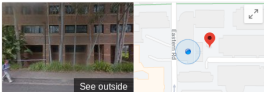
Macquarie University - Centre for Language Technology (CLT)
https://www.mq.edu.au/research/_technologies/_centre-for-language-technology-clt ▼
Centre for Language Technology. Located in Sydney, Australia, Macquarie University's Centre for Language Technology is Australasia's largest and longest-established body of researchers working in natural language processing, computational linguistics and language technology. We have a well-developed infrastructure ...

Macquarie University - What is language technology?
https://www.mq.edu.au/_data/assets/pdf_file/0008/_jmu_lt_program_2002.pdf ▼
Language Technology (LT) is the late 1990s outgrowth of 40 years of research into natural language processing (NLP), a subfield of artificial intelligence.

What is Language Technology? - Macquarie University
https://www.mq.edu.au/_data/assets/pdf_file/0008/_jmu_lt_program_2002.pdf ▼
The Language Technologies Institute at Carnegie Mellon educates the leaders of tomorrow and performs groundbreaking research in the areas of Natural Language Processing, Computational Linguistics, Information Extraction, Summarization & Question Answering, Information Retrieval, Text Mining & Analytics, ...

Language Technologies Institute - Carnegie Mellon University
<https://www.lti.cs.cmu.edu/> ▼
The Language Technologies Institute at Carnegie Mellon educates the leaders of tomorrow and performs groundbreaking research in the areas of Natural Language Processing, Computational Linguistics, Information Extraction, Summarization & Question Answering, Information Retrieval, Text Mining & Analytics, ...

DFKI LT - What is Language Technology?
<https://www.dfki.de/lt-general.php> ▼
Language technology — sometimes also referred to as human language technology — comprises ...


See outside

Centre for Language Technology ★
5.0 ★★★★★ 1 Google review
Language school in Macquarie Park, New South Wales
[Website](#) [Directions](#)

Address: Eastern Rd, Macquarie University NSW 2109
Phone: (02) 9850 7111
[Suggest an edit](#) · [Own this business?](#)

Add missing information
[Add business hours](#)

Know this place? [Answer quick questions](#)

Questions & answers
[Be the first to ask a question](#) [Ask a question](#)

Send to your phone [Send](#)

Reviews
1 Google review [Write a review](#) [Add a photo](#)

Google Search (13 Feb 2018)

The screenshot shows a Google search interface. The address bar displays the URL <https://www.google.com.au/search?client=>. The search bar contains the query "what's the best treatment for headaches?". Below the search bar, navigation tabs include "All", "Shopping", "Images", "Videos", "News", "More", "Settings", and "Tools". The results section indicates "About 3,890,000 results (0.69 seconds)".

The first search result is an advertisement from www.amrita.net/ titled "Headache Roll On Relief | Naturally Soothe Headaches | amrita.net". The ad text states: "Provides an Alternative to Conventional Treatment. Best Products and Prices." Below the text are links for "Facial Serums", "Essential Oils", "Shop Books", and "Natural Perfumes".

The second search result is an advertisement from www.healthnow.io/headache titled "Need Headache Information? | Call a HealthNow Doctor Today". The ad text includes: "Discuss Your Symptoms & Get Advice from an Expert from Just \$69. Call Now. Accredited & Registered · Australian Based GPs · Qualified After Hours GPs · Fully Accredited Doctors. Highlights: Provides Knowledgeable Medical Advice, Qualified Australian Doctors. How It Works · Meet The Team · What We Treat · Things To Know".

A box titled "Treatment may include:" contains a list of suggestions:

- Rest in a quiet, dark room.
- Hot or cold compresses to your head or neck.
- Massage and small amounts of caffeine.
- Over-the-counter medications such as ibuprofen (Advil, Motrin IB, others), acetaminophen (Tylenol, others), and aspirin.

Below the list is a link for "More items...".

Programme

1 Analysing Unstructured Data

2 Analysing Text Data

- Characteristics of Text
- Common Building Blocks for Text Analytics
- Some APIs for Text Analytics

Text as Arbitrary Symbols

- Words are encoded as arbitrary symbols.
- Different languages use different representations to represent the same word.
- Even within one language there is no clear correspondence between a word symbol and its meaning.



<https://www.linguisticsociety.org/content/how-many-languages-are-there-world>

Ambiguity everywhere I

Language features ambiguity at multiple levels.

Lexical

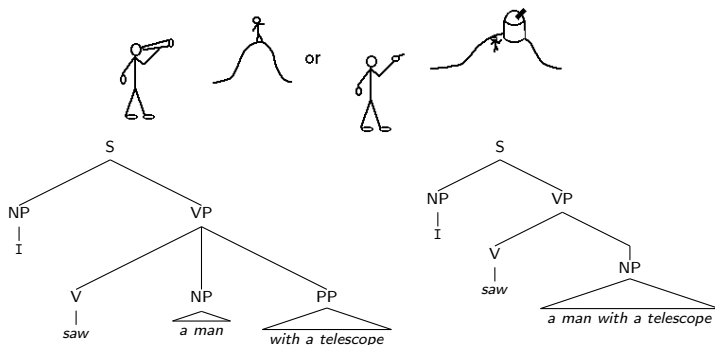
Example from Google's dictionary:

- bank (n): the land alongside or sloping down a river or lake.
- bank (n): financial establishment that uses money deposited by customers for investment, ...
- bank (v): form in to a mass or mound.
- bank (v): build (a road, railway, or sports track) higher at the outer edge of a bend to facilitate fast cornering.
- ...

Ambiguity everywhere II

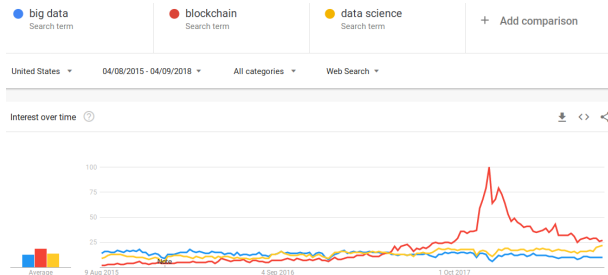
Syntactic

- “I saw a man with a telescope” ... who has the telescope?



So many words!

- Any language features a large number of distinct words.
- New words are coined.
- Words change their use in time.
- There are also names, numbers, dates... the possibilities are infinite.



<https://trends.google.com>

Programme

1 Analysing Unstructured Data

2 Analysing Text Data

- Characteristics of Text
- Common Building Blocks for Text Analytics
- Some APIs for Text Analytics

Tokenisation

- **Tokenisation**: Break down the input into words and other kinds of tokens.
- **Sentence Segmentation**: Break down the input into sentences.
- Tokenisation needs to be done as a first step in other applications.
- Same process as identifying separate units in programming languages, but harder.
- Tokenisation in space-delimited languages is fairly easy but some languages have no clear-cut way to separate words, or even sentences.

Keyword Extraction and Word Clouds

- **Keyword extraction:** Extract the most important words in a document or collection of documents.
- **Word cloud:** a graphical interface that displays words according to their importance.

How to Select and Score words?

- Remove stop words.
- Select words by frequency.
- Use tf.idf
- ...



Removing Stop Words

- Many packages offer lists of stop words.
- These lists include words that usually are not important.
- There is no universal list of stop words.

Stop words in the Python NLTK package

```
>>> from nltk.corpus import stopwords
>>> stopwords.words('english')
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which',
'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are',
'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do',
'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because',
'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against',
'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below',
'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again',
'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all',
'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't',
'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll',
'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn',
"didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven',
"haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't"]
```



Selecting Words by Frequency

- If you want to find words that **discriminate** between different documents ...
 - Very frequent words are not useful (because they are in most documents).
 - Very rare words are not useful (because they are in too few documents).
 - The right solution is somewhere in the middle.
- A practical solution is to apply this sequence:
 - 1 Remove stop words.
 - 2 Select the most frequent remaining words.

Selecting Words by $tf.idf$

- $tf.idf$ strikes a balance between words that are frequent but are not too frequent.
- **tf**: Term frequency. Words that are very frequent are more important.

$tf(w, d)$ = number of times word w occurs in document d

- **idf**: Inverse document frequency. Words that occur in many documents are less important.

$$idf(w) = 1 + \log\left(\frac{\text{number of documents}}{\text{number of documents containing word } w}\right)$$

- $tf.idf(w, d) = tf(w, d) \times idf(w)$
- We select words from document d with high $tf.idf$, possibly after removing stop words.

Stemming and Lemmatisation

- Words in many languages (e.g. English) have inflections.
 - Singular, plural, verb-ing, etc.
- Stemming and lemmatisation allow to group words that are different only because of their inflections.
- **Stemming**: Remove the part of a word that has the inflection to produce the **stem**.
- **Lemmatisation**: Convert an inflected word into a word without inflections to produce the **lemma** or **base form**.
- Stemming is easier and requires less knowledge of the language. Often stemming is all you need.
- Lemmatisation is useful when you want to produce real words.
 - E.g. if you want to display keywords.

Part of Speech Tagging

- Words with the same part of speech have similar grammatical properties.
- In general, one can replace a word with another of the same part of speech and the sentence is still grammatical.
- Most words belong to **open class types**: nouns, verbs, adjectives, adverbs.
 - These words usually determine the topic of the sentence.
 - For example, keywords would normally be words in open class types.
- Words in the **closed class types** are useful to connect other words: prepositions, determiners, pronouns, conjunctions,
 - These words are usually removed by some text applications.
 - For example, stop words are normally words from closed class types.

Parts of Speech in the Penn Treebank

CC	Coordinating conjunction	PRP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP\$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Parts of Speech in SAS Text Miner

Abbr (abbreviation)

Adj (adjective)

Adv (adverb)

Aux (auxiliary or modal)

Conj (conjunction)

Det (determiner)

Interj (interjection)

Noun (noun)

Num (number or numeric
expression)

Part (infinitive marker,
negative participle, or
possessive marker)

Pref (prefix)

Prep (preposition)

Pron (pronoun)

Prop (proper noun)

Punct (punctuation)

Verb (verb)

VerbAdj (verb adjective)

Named Entity Recognition

- **Named entities** are (often multi-word) expressions that refer to proper names of specific types.
 - Persons, organisations, locations, artifacts, dates, . . .
- Named entity recognition is one of the most common tasks in text analytics.

When **Sebastian Thrun** PERSON started working on self-driving cars at **Google** ORG in **2007** DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major **American** NORP car companies would shake my hand and turn away because I wasn't worth talking to," said **Thrun** PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** ORG **earlier this week** DATE.

A little **less than a decade later** DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

<https://explosion.ai/demos/displacy-ent>

Entities in the Message Understanding Conference

- Named Entities
 - Organization
 - Person
 - Location
- Temporal Expressions
 - Date
 - Time
- Number Expressions
 - Money
 - Percent

MUC

- Initiated and financed by DARPA (Defense Advanced Research Projects Agency).
- From 1987 to 1997.
- The goal was to advance methods for information extraction from text.
- MUC-6 (1995) introduced the task of named entity recognition.
- The MUC named entities have been used by many systems since then.

Entities in SAS Text Miner

ADDRESS	(postal address or number and street name)
COMPANY	(companyname)
CURRENCY	(currencyorcurrencyexpression)
DATE	(date, day, month, or year)
INTERNET	(email address or URL)
LOCATION	(city, country, state, geographical place/region, political place/region)
MEASURE	(measurement or measurement expression)
ORGANIZATION	(government, legal, or service agency)
PERCENT	(percentage or percentage expression)
PERSON	(person's name)
PHONE	(phone number)
PROP_MISC	(proper noun with an ambiguous classification)
SSN	(Social Security number)
TIME	(time or time expression)
TIME_PERIOD	(measure of time expressions)
TITLE	(person's title or position)
VEHICLE	(motor vehicle including color, year, make, and model)

Text Classification

Many different tasks can be seen as text classification.

- E-mail filtering, spam detection, sentiment analysis ...

To classify text it needs to be converted into a vector of features.

Feature Selection

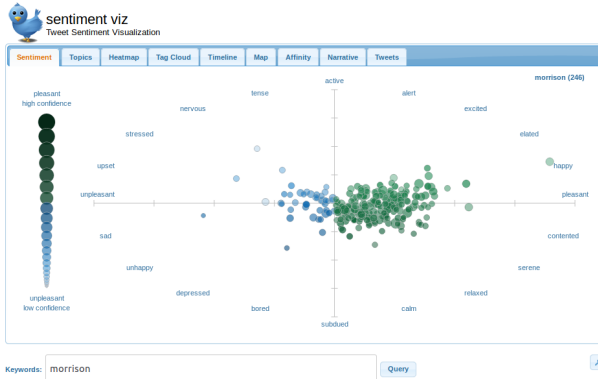
- Extract key words and use them to build document vectors for classification.
- For example, remove **stop words** and/or select words with high tf.idf.

Feature Extraction

- Generate document vectors based on mathematical and statistical combinations of the entire information of the text.
- **Latent Semantic Analysis (LSA)**, **Singular Value Decomposition (SVD)** and **Principal Component Analysis (PCA)** are traditionally used for feature extraction.
- More recent approaches use **neural networks** and **word embeddings**.

Sentiment Analysis

- Sentiment analysis is a popular example of text classification.
- Often needed for market analysis.
- A well known approach to analyse social media posts.



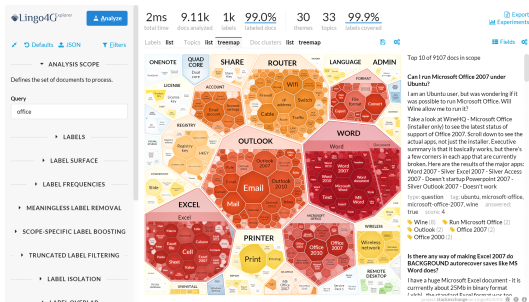
https://www.csc2.ncsu.edu/faculty/healey/tweet-viz/tweet_app/

Text Retrieval / Filtering

- Often needed to find specific information in large volumes of text.
- Search engines are the first popular applications of text retrieval.
- A common step before doing other processing tasks such as sentiment analysis.

Text Clustering

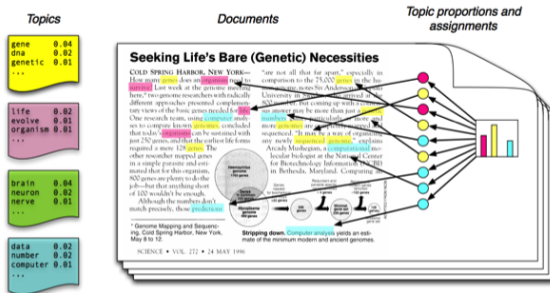
- Nothing to do with computer clusters ...
- Useful when we have large volumes of text but no labels.
- Can help characterise types of customers, common views of opinion, etc.



<https://get.carrotsearch.com/lingo4g/1.4.0/doc/>

Topic Modelling

- Topic modelling is a more complex form of unsupervised text processing.
- The task is to find the common topics in a collection of texts (e.g. tweets).
- Topic modelling often returns keywords that are most characteristic of each topic.



Programme

1 Analysing Unstructured Data

2 Analysing Text Data

- Characteristics of Text
- Common Building Blocks for Text Analytics
- Some APIs for Text Analytics

Web Demos

- <https://explosion.ai/demos/>
- https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
- <https://developer.aylien.com/text-api-demo>
- <http://text-processing.com/demo/>
- ...

Programming Libraries

- Spacy <https://spacy.io/>
- Natural Language Toolkit (NLTK) <https://www.nltk.org/>
- Scikit-Learn <http://scikit-learn.org/stable/>
- Keras <https://keras.io/>
- ...

Graphical Interfaces

Usually integrated in general machine learning tools

- RapidMiner
- Weka
- SAS Enterprise Miner
- ...

Cloud Services

- Azure Machine Learning Studio
<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/text-analytics>
- Aylien <https://aylien.com/text-api/>
- ...

Comparison for Named Entity Recognition

Text APIs compared in these two posts:

- <https://medium.com/@boab.dale/text-analytics-apis-part-1-the-bigger-players-3ce8a93577bd>
- <https://becominghuman.ai/text-analytics-apis-part-2-the-smaller-players-c9e608cf7810>

Table 2. *Results on the CoNLL shared task data; all values are percentages*

	Amazon comprehend			Google NL			IBM NL		
	Prec'n	Recall	$F_{\beta=1}$	Prec'n	Recall	$F_{\beta=1}$	Prec'n	Recall	$F_{\beta=1}$
LOC	76.13	72.66	74.36	58.81	86.45	70.00	70.17	86.15	77.34
MISC	58.40	10.40	17.65	36.76	19.37	25.37	2.08	0.14	0.27
ORG	74.72	59.24	66.08	68.03	48.16	56.40	69.86	27.63	39.60
PER	87.14	82.99	85.02	82.45	83.36	82.90	73.13	76.07	74.57
Overall	78.95	63.93	70.65	66.15	65.97	66.06	70.51	55.36	62.03

<https://medium.com/@boab.dale/text-analytics-apis-part-1-the-bigger-players-3ce8a93577bd>

Take-home Messages

- Sources of unstructured data.
- Impact of unstructured data.
- Characteristics of text.
- Common building blocks for text analytics.

What's Next

Week 9

- Visual Analytics.
- Assignment 3 released.