Advanced Data & Network Mining

Introduction to Text Mining

2023-24

terri.hoare@dbs.ie

Data Mining
Unstructured Data : Text Mining

Eric Siegel in his book **Predictive Analytics** (Siegel, 2013) provides an interesting analogy, if all the data in the world was equivalent to the water on earth, then textual data is like the ocean, making up the majority of the volume.

Some of the first applications of text mining came about when people were trying to categorise documents [Cutting, 1992] and [Hearst, 1999] recognised that text analysis does not require artificial intelligence but a “mixture of computationally driven and user guided analysis” which is at the heart of the supervised models used in predictive analytics.

Text Mining, more than any other type of data mining fits the ‘mining’ metaphor as we attempt to separate valuable keywords from a mass of other words (or relevant documents from a sea of documents) and use them to identify meaningful patterns or make predictions.

Text Mining : Case Study
"It is NLP my dear Watson!"

Perhaps the most famous application of text mining is IBM's Watson program, which performed spectacularly when competing against humans on the nightly game show Jeopardy! How does Watson use text mining? Watson has instant access to hundreds of millions of structured and unstructured documents, including the full content of Wikipedia entries.

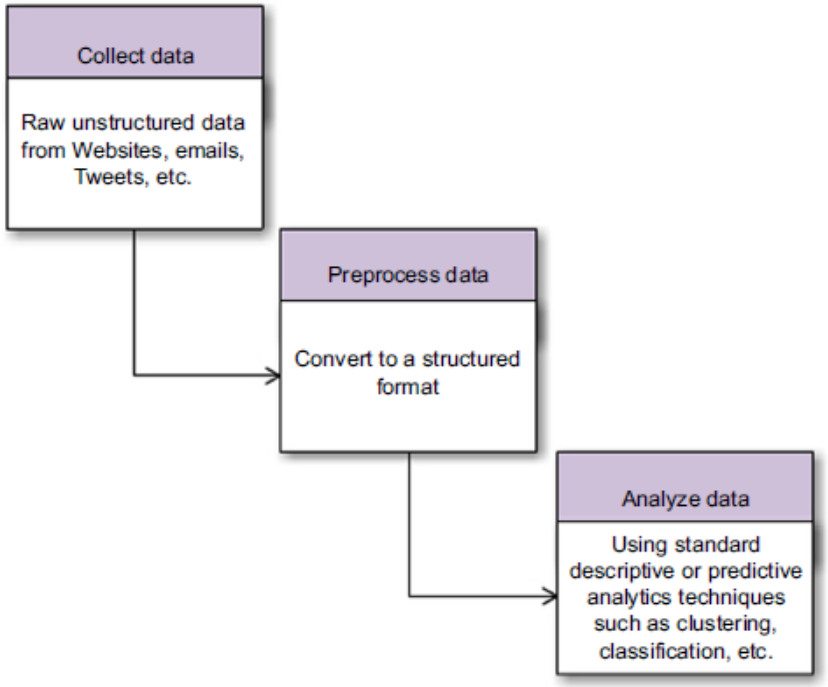
When a Jeopardy! question is transcribed to Watson, it searches for and identifies candidate documents that score a very close match to the words of the question. The search and comparison methods it uses are similar to those used by search engines, and include many of the techniques, such as n-grams and stemming, which we discuss in this chapter. Once it identifies candidate documents, it again uses other text mining (also known as natural language processing or NLP) methods to rank them. For example, if the answer is, REGARDING THIS DEVICE, ARCHIMEDES SAID, "GIVE ME A PLACE TO STAND ON, AND I WILL MOVE THE EARTH, a Watson search for this sentence in its databases might reveal among its candidate documents several with the term "lever." Watson might insert the word "lever" inside the answer text and rerun a new search to see if there are other documents with the new combination of terms. If the search result has many matches to the terms in the sentence—as it most likely would in this case—a high score is assigned to the inserted term.

If a broad and non-domain-focused program like Watson, which relies heavily on text mining and NLP, can answer open-ended quiz show questions with nearly 100% accuracy, one can imagine how successful specialized NLP tools would be. In fact IBM has successfully deployed a Watson-type program to help in decision making at health care centers (Upbin, 2013).

Text mining also finds applications in numerous business activities such as email spam filtering, consumer sentiment analysis, and patent mining to name a few. We will explore a couple of these in this chapter.

Data Mining
Text Mining : How it Works

The three main steps in text Mining :-



A high-level process for text mining.

Data Mining
Text Mining : Pre-processing : TF-IDF

Consider a web search problem where the user types in some keywords and the search engine extracts all the documents (web pages) that contain these keywords.
Search Engines apply the logic:-

1. Give a high weightage to those keywords that are relatively rare **TF**

TF = $\frac{n_k}{n}$ (ratio of number of times keyword appears in document to total number terms in the document)

2. Give a high weightage to those webpages that contain a large number of instances of “rare” keywords **IDF**

IDF = $\log_2(\frac{N}{N_k})$ where N is the total number mined documents and N_k is the number of documents that contain the keyword k

3. **TF-IDF =** $\frac{n_k}{n} * log_2(\frac{N}{N_k})$

Data Mining
Text Mining : Pre-processing : TF-IDF cont.

The highest weighted web pages are those with the highest **TF-IDF** score that is highest product of TF (Term Frequency) and IDF (Inverse Document Frequency)

Therefore, only those pages that not only contain the rare keywords but have a high number of instances of those keywords should appear at the top of the search results

Typically **TF-IDF** scores for every word in the set of documents is created in the pre-processing step of the three step process

Data Mining
Text Mining : Pre-processing : Terminology

Consider two sentences (either in an email, in two different text files or in the same text file) for which the objective is to do similarity map:-

1. “This is a book on data mining”
2. “This book describes text mining and data mining using RapidMiner

Tokenization is the process of discretising words (**tokens**) in documents (above each sentence is a document)

A **TDM** (Term Document Matrix) or document vector is created where each document is an example (row) and each token is an attribute (col).

Bag-of-words technique, the following can be used to weight the tokens: -

- term counts
- term frequencies (TF)
- Term frequency-inverse document frequency (TF-IDF) – benchmark
- binary flags

Data Mining
Text Mining : Pre-processing : Terminology

Document Vector or Term Document Matrix…

Building a Matrix of Terms from Unstructured Raw Text

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	using
Document 1	1	1	1	1	1	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	2	1	1	1	1	1

Using Term Frequencies Instead of Term Counts in a TDM

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	Using
Document 1	1/7 = 0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0.1428	0	0	0	0	0
Document 2	1/10 = 0.1	0	0	0.1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1

ExampleSet (2 examples, 0 special attributes, 12 regular attributes)

Row No.	RapidMiner	This	a	and	book	data	describes	is	mining	on	text	using
1	0	0	0.577	0	0	0	0	0.577	0	0.577	0	0
2	0.447	0	0	0.447	0	0	0.447	0	0	0	0.447	0.447

Calculating TF–IDF scores for the sample TDM.

Data Mining
Text Mining : Pre-processing : Terminology cont.

Stopword filtering is applied to remove common words such as ‘a’, ‘this’, ‘and’

Row No.	RapidMiner	book	data	describes	mining	text	using
1	0	1	1	0	1	0	0
2	1	1	1	1	2	1	1

Stopword filtering reduces the size of the TDM significantly.

Data Mining
Text Mining : Pre-processing : Terminology cont.

Term Filtering and **Lexical Substitution**

- Removal of common industry terms for example ‘car’ in the automotive industry.
- Substitution from a dictionary for terms meaning the same.

Stemming

- Most common in English is the Porter Method (1980) which standardises suffixes for example replacing suffixes of terms ending ‘ies’ with ‘y’
- Language and period sensitive for example Shakespeare versus present-day literature

Data Mining
Text Mining : Pre-processing : Terminology cont.

N-Grams are families of words in spoken and written language for example ‘Good’ followed by ‘Morning’, ‘Night’, ‘Day’.

Google has processed more than a trillion (1,04,908,267,229) words back as far as 2006 and published the counts for all (1,176,470,663) five word (5-gram) sequences that appear at least 40 times [Franz, 2006]. In practice bigrams and trigrams are useful providing they are not too computationally expensive or large to build

Row...	label	RapidMiner	book	book_data	book_descr...	data	data_mining	describes	describes_data	mining	mining_text	mining_usi...	text_0	text_mining	using	using_RapidMiner
1	text1	0	0.447	0.447	0	0.447	0.447	0	0	0.447	0	0	0	0	0	0
2	text2	0.243	0.243	0	0.243	0.243	0.243	0.243	0.243	0.485	0.243	0.243	0.243	0.243	0.243	0.243

Meaningful n-grams show higher TF–IDF scores.

Data Mining
Text Mining : Pre-processing : Steps

Pre-processing steps in summary:-

A Typical Sequence of Preprocessing Steps to Use in Text Mining

Step	Action	Result
1	Tokenize	Convert each word or term in a document into a distinct attribute
2	Stopword removal	Remove highly common grammatical tokens/words
3	Filtering	Remove other very common tokens
4	Stemming	Trim each token to its most essential minimum
5	n-grams	Combine commonly occurring token pairs or tuples (more than 2)

Data Mining
Text Mining : Implementation

Test Mining with Clustering and Classification – two Case Studies:-

- **Case Study 1**

Group keywords found on several web pages (documents) using clustering techniques. This simple example can be easily extended to a more comprehensive document-clustering problem where the most common words occurring in a document are used as flags to group multiple documents

- **Case Study 2**

Classify Blogs by Gender using a set of blogs (documents) that have been gender-author labelled to train the model

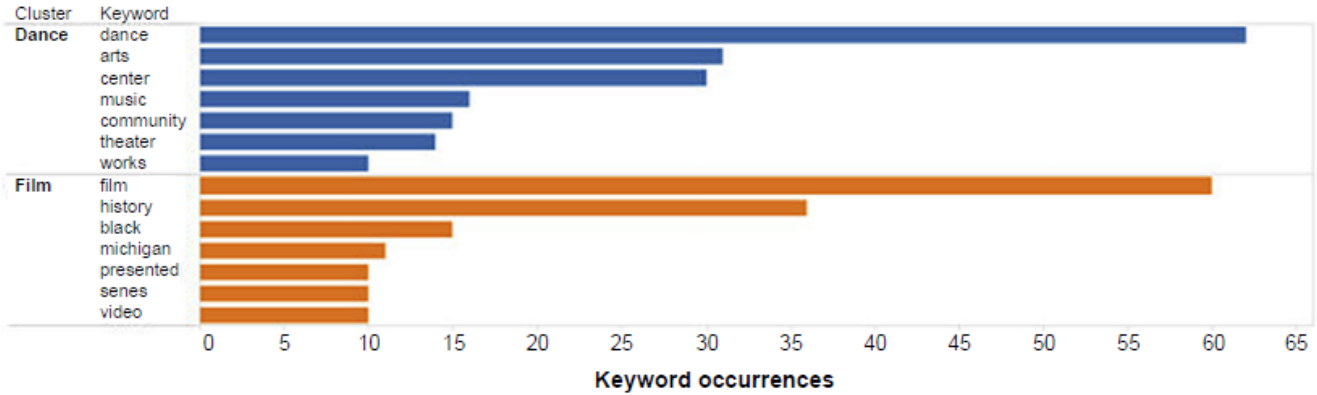
Data Mining
Text Mining : Case Study 1 : Keyword Clustering

The site (http://www.detroitperforms.org) is hosted by a public television station and is meant to be used as a platform for reaching out to members of the local community who are interested in the arts and culture. The aim is both to reach out as well as to eventually aid in targeted marketing campaigns meant to attract donors to public broadcasting.

The site has pages for several related categories: Music, Dance, Theatre, Film and so on. Each of these pages contains articles and events related to that category.

The goal is to characterise each page on the site and identify the top keywords that appear on each page by crawling each category page, extracting content, and converting into a structured document vector consisting of keywords. Finally using k-medoids clustering to sort and rank the keywords. (k-medoids clusters around the most centrally located object in cluster. Similar k-means, not as susceptible to noise and outliers).

Data Mining
Text Mining : Case Study 1 : Keyword Clustering cont.



Results of the website keyword clustering process.

Text Mining
Case Study 2 : Predicting Blog Author Gender

The objective is to attempt to predict the gender of Blog authors based on the content of the Blog

RapidMiner demonstration files as below. Note that the text extension will need to be installed from the RapidMiner Help Menu

09_Text_9.3.2_blog_gender_step2_preprocess.rmp
09_Text_9.3.2_blog_gender_step3.1_key_features.rmp
09_Text_9.3.2_blog_gender_step3.2_build_predictive_models.rmp
09_Text_9.3.2_blog_gender_step4.1_docVec_testing.rmp
09_Text_9.3.2_blog_gender_step4.2_apply_model_test_data.rmp
09_Text_9.3.2_blog-gender-dataset.xlsx
09_Text_9.3.2_blog-gender-dataset-removed-missing.xlsx
09_Text_9.3.2_testing-sample.xlsx

Text Mining
Case Study 2 : Predicting Blog Author Gender cont.

- **Step 1 – Gather Unstructured Data**

Data set consists of more than 3000 individual blogs from men and women around the world as below. Raw data for the Case Study will be split 50:50 for training and testing of the performance of the model purposes

Raw Data for the Blog Classification Study

BLOG	GENDER
This game was a blast. You (as Drake) start the game waking up in a train that is dangling over the side of a cliff. You have to climb up the train car, which is slowly teetering off the edge of the cliff, ready to plummet miles down into a snowy abyss. From the snowy beginning there are flashbacks to what led Drake to this predicament. The story unfolds in a very cinematic manner, and the scenes in between levels, while a bit clichéd by Hollywood standards, are still just as good if not better than your average brainless Mel Gibson or Bruce Willis action movie. In fact, the cheese is part of the fun and I would venture to say it's intentional.	M
My mother was a contrarian, she was. For instance, she always wore orange on St. Patrick's Day, something that I of course did not understand at the time, nor, come to think of it do I understand today. Protestants wear orange in Ireland, not here, but I'm pretty sure my mother had nothing against the Catholics, so why did she do it? Maybe it had to do with the myth about Patrick driving the snakes, a.k.a. pagans, out of Ireland. Or maybe it was something political. I have no idea and since my mother is long gone from this earth, I guess I'll never know.	F
LaLicious Sugar Soufflé body scrub has a devoted following and I now understand why. I received a sample of this body scrub in Tahitian Flower and after one shower with this tub of sugary goodness, I was hooked. The lush scent is deliciously intoxicating and it ended up inspiring compliments and extended sniffing from both loved ones and strangers alike. Furthermore, this scrub packs one heck of a punch when it comes to pampering dry skin. In fact, LaLicious promises that this body scrub is so rich that it will eliminate the need for applying your post-shower lotion. This claim is true — if you follow the directions.	F
Stopped by the post office this morning to pick up a package on my way to the lab. I thought it would be as good a time as any to clean up my desk and at the very least make it appear that I am more organized than I really am (seriously, it's a mess). It's pretty nice here on the weekends, it's quiet, there's less worry of disturbing undergrad classes if I do any experiments in the daytime.	M

Text Mining
Case Study 2 : Predicting Blog Author Gender cont.

Learn Example Process Overview

- **09_Text_9.3.2_blog_gender_step2_preprocess.rmp**

Remove Records Missing Values
Split Data (50:50 Train:Test)
Build WordList (Occurrences:-Total; Document; M; F)
Build Document Vector (Each Blog with words as attributes TF-IDF)

- **09_Text_9.3.2_blog_gender_step3.1_key_features.rmp**

Weight Attributes for selection using Information Gain and SVM

- **09_Text_9.3.2_blog_gender_step3.2_build_predictive_models.rmp**

Modelling – select attributes by weight, train using different models

- **09_Text_9.3.2_blog_gender_step4.1_docVec_testing.rmp**

Prepare Test Set

- **09_Text_9.3.2_blog_gender_step4.2_apply_model_test_data.rmp**

Apply Model to Test Set

Data Mining
Text Mining : Discussion Points

- Unstructured data of which text is a major portion appears to be doubling in volume every year, The ability to be able to automatically process and mine data from such digital data will become an important skill for the future

- Introduced a three-step process and key tools such as tokenisation, stemming, n-gramming, and stopword removal

- Unstructured data can be mined using the same algorithms as for other data. The key is to convert the unstructured data into a semi-structured format

- Concepts like TF-IDF allow us to transform a corpus of text into a matrix of numbers which can be worked on by the standard machine learning algorithms

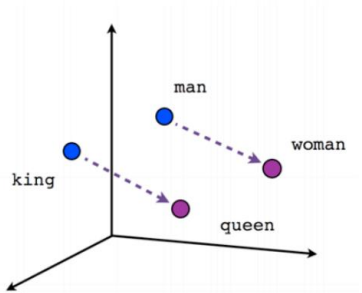
Data Mining
Text Mining : Discussion Points

We have considered a bag-of-words approach to vectorise the unstructured text. There are also embedding techniques such as word-2-vec; doc-2-vec; BERT; GloVe for example that better capture context.

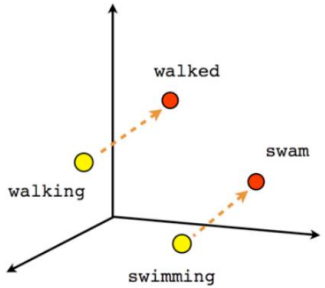
Either pre-trained or generated word embeddings can be used.

- Training word vector embeddings across very large datasets takes enormous computation which is avoided by pre-computed embeddings
- Pre-computed embeddings may be out of date and may not suit specialised vocabulary
- Using a pre-trained word-embedding can amplify a smaller dataset for an ML task
- Word embeddings have become an essential component for modern NLP pipelines
- Word2vec produces an embedding vector for each word from a large corpus of text, such that words with similar characteristics are close to each other in embedding space.

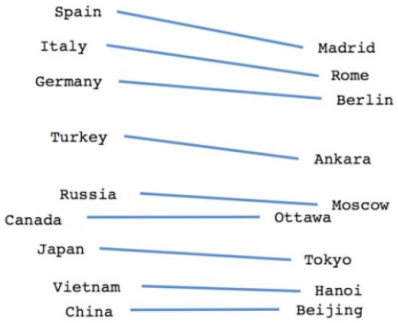
Text Analytics - NLP
Sequence Models – word embedding algebra



Male-Female



Verb tense



Country-Capital

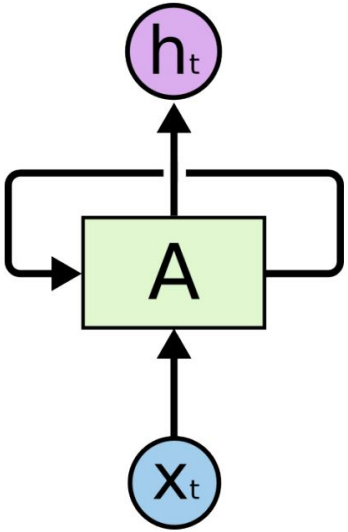
Deep Learning - RNN
State-of-the-art

THE SONNETS

by William Shakespeare

From fairest creatures we desire increase,
That thereby beauty's rose might never die,
But as the riper should by time decease,
His tender heir might bear his memory:
But thou, contracted to thine own bright eyes,
Feed'st thy light's flame with self-substantial fuel,
Making a famine where abundance lies,
Thyself thy foe, to thy sweet self too cruel:
Thou that art now the world's fresh ornament,
And only herald to the gaudy spring,
Within thine own bud buriest thy content,
And tender churl mak'st waste in niggarding:
    Pity the world, or else this glutton be,
    To eat the world's due, by the grave and thee.

When forty winters shall besiege thy brow,
And dig deep trenches in thy beauty's field,
Thy youth's proud livery so gazed on now,
Will be a tatter'd weed of small worth held:
Then being asked, where all thy beauty lies,
Where all the treasure of thy lusty days;
To say, within thine own deep sunken eyes,
Were an all-eating shame, and thriftless praise.
How much more praise deserv'd thy beauty's use,
If thou couldst answer 'This fair child of mine
Shall sum my count, and make my old excuse,'
Proving his beauty by succession thine!
    This were to be new made when thou art old,
    And see thy blood warm when thou feel'st it cold.



Text Analytics - NLP
Deep learning (RNN)

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

Text Analytics – NLP – State-of-the-art
Attention – Transformers

Memory is **attention** through time. Alex Graves 2020

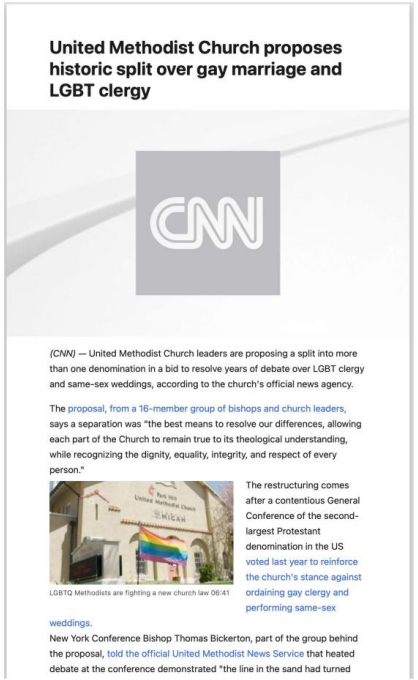
BERT
unsupervised language model
340m parameters
training: 4xdays on 64TPU's

GPT-2
autoregressive language model
1.5b parameters
training: 7xdays on 256TPU's

GPT-3
autoregressive language model
175b parameters
training: 12xdays on 10k+ TPU's cost~ $4.6m

(Autoregressive models take the previous predictions to generate a new prediction)

Text Analytics – NLP - State-of-the-art Attention – Transformers – GPT-3



United Methodist Church proposes historic split over gay marriage and LGBT clergy

CNN

(CNN) — United Methodist Church leaders are proposing a split into more than one denomination in a bid to resolve years of debate over LGBT clergy and same-sex weddings, according to the church's official news agency.

The proposal, from a 16-member group of bishops and church leaders, says a separation was "the best means to resolve our differences, allowing each part of the Church to remain true to its theological understanding, while recognizing the dignity, equality, integrity, and respect of every person."

LGBTQ Methodists are fighting a new church law 06:41

The restructuring comes after a contentious General Conference of the second-largest Protestant denomination in the US voted last year to reinforce the church's stance against ordaining gay clergy and performing same-sex weddings.

New York Conference Bishop Thomas Bickerton, part of the group behind the proposal, told the official United Methodist News Service that heated debate at the conference demonstrated "the line in the sand had turned

CNN

After two days of intense debate, the United Methodist Church has agreed to a historic split – one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination. The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

GPT-3

Text Analytics – NLP – State-of-the-art
NLP Tasks

NLP (Natural Language Processing) - a way for computers to analyse, understand, and derive meaning from human language

- Text Classification (for example sentiment; emotion; disease code)
- Text Categorisation (for example clustering; topic modelling)
- Named Entity Recognition (NER)
- Part-of-Speech Tagging
- Semantic Parsing and Question Answering
- Paraphrase Detection
- Language Generation
- Mult-document Summarisation
- Machine Translation
- Speech Recognition

References Text Mining and NLP

Bengfort, B., Bilbro, R, Ojeda, T. (2019) *Applied Text Analysis with Python*. Newton, MA: O'Reilly Media.

Blei, D., Ng, A., Jordan, M. (2003) 'Latent Dirichlet Allocation'. *Journal of Machine Learning Research*, 3, pp.993-1022. Available at: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (Accessed: 11 August 2023).

Jurafsky, D., Martin, J. (2023) *Speech and Language Processing.* 3rd edn. Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (Accessed: 2 August 2023).

Kerr, C., Hoare, T., Carroll, P., Marecek, J. (2020) 'Ensemble of Temporal Language Classifiers'. *Data Mining and Knowledge Discovery*, 34, pp. 532-562. Available at: https://link.springer.com/article/10.1007/s10618-019-00671-x (Accessed: 11 August 2023).

References Text Mining and NLP

Kotu, V. and Deshpande, B. (2019) *Data Science Concepts and Practice.* 2nd edn. Burlington, MA: Morgan Kaufmann Publishers.

Li, J., Chen, X., Hovy, E. and Jurafsky, D. (2015) 'Visualizing and understanding neural models'. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.681–691.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013a) 'Efficient Estimation of Word Representations in Vector Space'. Available at: https://www.researchgate.net/publication/234131319_Efficient_Estimation_of_Word_Representations_in_Vector_Space (Accessed: 11 August 2023).

References Text Mining and NLP

.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013b) 'Distributed representations of words and phrases and their compositionality'. *Advances in Neural Information Processing Systems*, 26, pp.1-9. Available at: https://www.researchgate.net/publication/257882504_Distributed_Representations_of_Words_and_Phrases_and_their_Compositionality (Accessed: 11 August 2023).

North, M. (2018) *Data Mining for the Masses.* 3rd edn. Global Text Project. Available at: https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf (Accessed: 26 June 2023).

Skiena, S. (2020) *The Algorithm Design Manual.* 3rd edn. Basel: Springer Nature Switzerland AG.

Weizenbaum, J. (1966) 'ELIZA – A computer program for the study of natural language communication between man and machine'. *Communications of the ACM*, 9(1), pp.36-45.

Welch, C. (2018) *Google just gave a stunning demo of Assistant making an actual phone call*. Available at: https://www.theverge.com/2018/5/8/17332070/google-assistant-makes-phone-call-demo-duplex-io-2018 (Accessed: 11 August 2023).

References Text Mining and NLP

Welch, C. (2018) *Google just gave a stunning demo of Assistant making an actual phone call*. Available at: https://www.theverge.com/2018/5/8/17332070/google-assistant-makes-phone-call-demo-duplex-io-2018 (Accessed: 11 August 2023).