STUTERN

Natural Langugage Processing (NLP)

SGA07_DATASCI

09th April 2020

Book Keeping

- Group Tasks (25% of total course score)
  - Due date 24th April
- Two more weeks for learning modules
  - Data Security & Gist: IoT Engineering
- 4 weeks for Final Project (50% of total course score)
  - Due date 29th May

Module Overview

- Overview of Text Processing
- Core Concepts: Bag of words, Tokenisation, Stemming, etc
- TFIDF, Topic Modelling & Sentiment Analysis
- Dialog Systems
- NLP in R

NLP (Def.)

- Interaction between data science and human language
- Conversations in the form of unstructured data
- Application in Healthcare, Personal Assistants, Search Engines, Translators, etc

> Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages like speech or text.

Text Preprocessing

- Regular Expression
- From Caps to lower cases
- Split into words or phrases
- Special characters
- Replace characters, words or expressions
- Extract numbers and non-texts

Core Concepts

Bag of words is the process of counting all the words in a piece of text

Tokenisation is the task of cutting a text into pieces called tokens, and at the same time throwing away certain characters, such as punctuation

Stop word removal is the process of getting rid of common language articles, pronouns and prepositions such as “and”, “the” or “to” in English

Stemming Refers to the process of slicing the end or the beginning of words with the intention of removing affixes (lexical additions to the root of the word)

Lemmatization Has the objective of reducing a word to its base form and grouping together different forms of the same word

TF-IDF

- Term Frequency - Inverse Document Frequency
- Measure of how important a word may be in terms of its frequency of occurrence
- Decrease the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

> The statistic tf-idf is intended to measure how important a word is to a document in a collection (or corpus) of documents, for example, to one novel in a collection of novels or to one website in a collection of websites.

TF-IDF

- Term Frequency - Inverse Document Frequency
- Measure of how important a word may be in terms of its frequency of occurrence
- Decrease the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents
- give us insight into how language is used in a collection of natural language

$$idf(term) = In(\frac{n_{documents}}{n_{documents\_containing\_term}})$$

Practice Lab

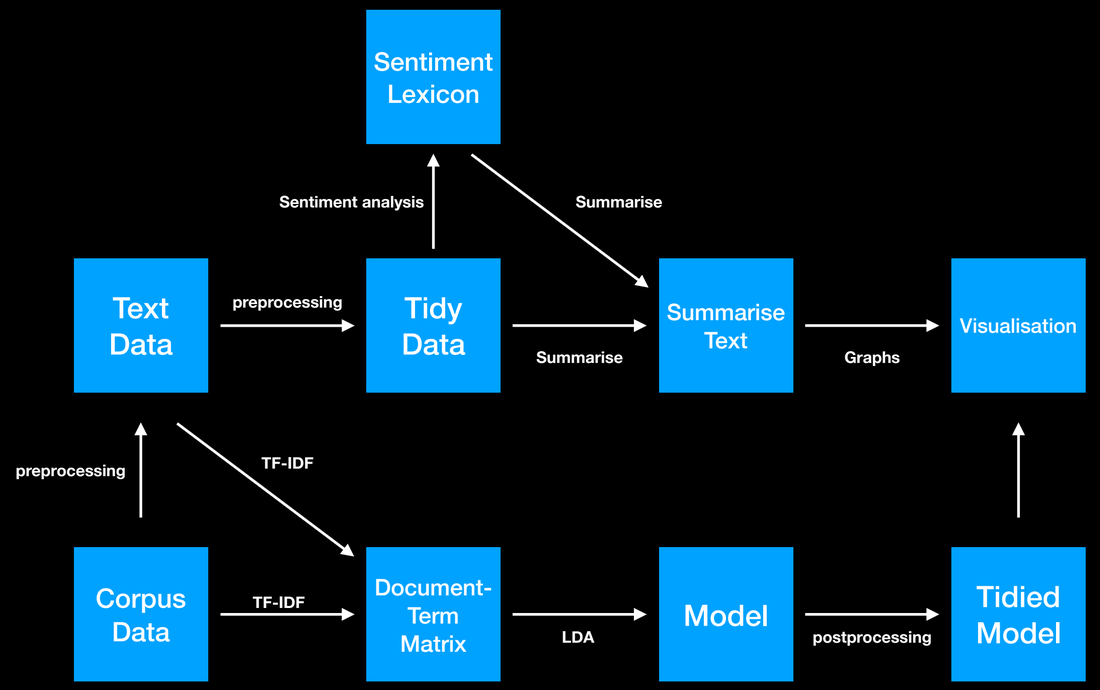
Analyse word and document frequency: tf-idf

Use the following Instructions:

- Get your data in R
- Explore distribution of terms
- Explore inverse proportionality of terms and ranks
- Bind tf-idf function using “tidytext” package

STUTERN

Topic Modelling



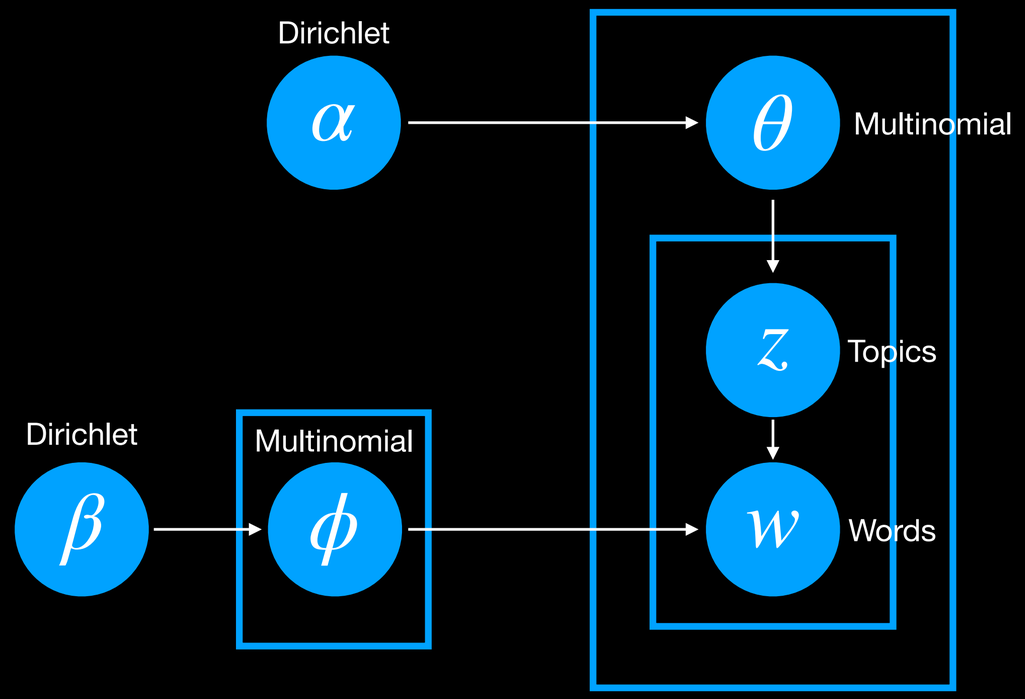
> Topic modelling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items

Latent Dirichlet Allocation

- Geometric approach to text classification
- Word-topic probabilities
- Document-topic probabilities

> LDA is a mathematical method for estimating both of these at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document.

Latent Dirichlet Allocation



$$P(W, Z, \theta, \phi; \alpha, \beta) = \underbrace{\Pi_{j=1}^{M} P(\theta_j; \alpha) \Pi_{i=1}^{K} P(\phi_i; \beta)}_{\text{Dirichlet Distribution}} \underbrace{\Pi_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \phi_j z_{j,t})}_{\text{Multinomial Distribution}}$$

Practice Lab

Build a Topic Model in R

Use the following Instructions:

- Get your data in R
- Use the LDA function of the “topicmodels” package
- Explore word-topic probabilities
- Explore document-topic probabilities

Sentiment Analysis

- Opinion Mining | Sentiment Lexicons
- Sentiment of a text is to consider the text as a combination of its individual words
- Sentiment content of the whole text as the sum of the sentiment content of the individual words

> Text mining approach to understand the emotional intent of words to infer whether a section of text is positive or negative, or perhaps characterised by some other more nuanced emotion like surprise or disgust.

Practice Lab

Build a Sentiment Analysis in R

Use the following Instructions:

- Get your data in R
- Convert to tidy data
- Filter for sentiments and give score
- Visualise your sentiment score
- Use word cloud to visualise high ranked positive sentiment words

Dialog Systems (Def.)

- Task-oriented: designed for a particular task and set up to have short conversations (from as little as a single interaction to perhaps half-a-dozen interactions) to get information from the user to help complete the task
- Chatbot: designed for extended conversations, set up to mimic the unstructured conversational or 'chats' characteristic of human-human interaction, rather than focused on a particular task
- Examples: Siri, Contana, Alexa

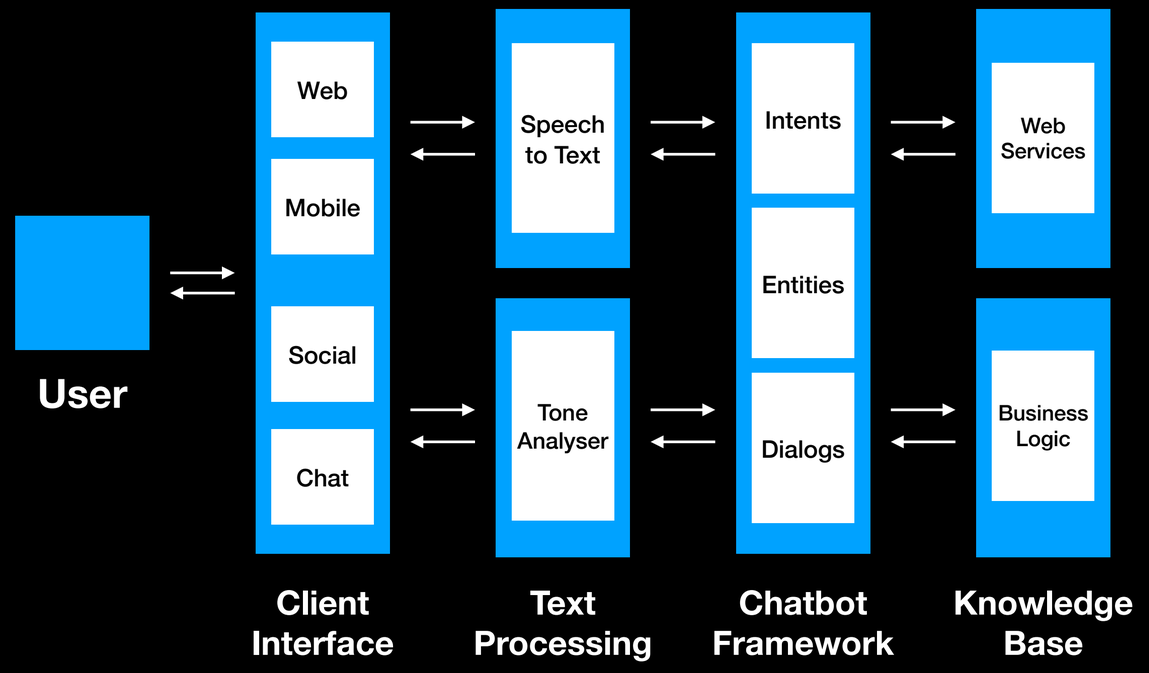
> "Dialog systems or conversational agents are programs that communicate with users in natural language (text, speech, or even both) to mimic sentience in humanity."

Properties of Dialog Systems

- Turn-taking: Questions and answers made up of single word or multiple sentences
- Conversational Implication: To hold and pass around contextual inference
- Grice's maxims: Quality, quantity, relevance and manner of text/speech
- Speech acts: Assertiveness, Directives, Commissions, Expressions & Declarations
- Prosody: Rhythm, intonation, stress & sentiments (emotions)

Chatbot Architecture

- Rule-based
- Corpus-based
  - Information-retrieval
  - Machine learned sequence transduction



Practice Lab

Build a Sentiment Analysis for Covid-19 Tweets in R

Use the following Instructions:

- Get your data in R
- Preprocess to clean data (text)
- Filter for sentiments and give score
- Visualise your sentiment score

Next Steps

Try to move your model into production

Use the following Instructions:

- Reproduce the R code in python
- Build a simple web app with flask & Bootstrap | Build with Shiny App in R
- Frontend interface should collect user email, frequency of update, word cloud, latest tweets & sentiment graph
  - Explore graph that shows sentiments over time
- Improve the "Get tweets" to update every hour (or based on frequency of update specified)
  - Beware: you might want to join old and new data
  - Inform: send a summary email on the pandemic state to user email
- Write a function to alert user by email when the sentiment changes from negative to positive

Recap/Summary

At the end of this Module, you should understand;

- Basic concepts and commands for text processing
- Using term frequency and inverse document frequency allows us to find words that are characteristic for one document within a collection of documents
- Introduces topic modelling for finding clusters of words that characterise a set of documents
- Introduced how to approach sentiment analysis
- General overview of dialog systems, properties and architecture
- Implemented sentimental analysis for Covid-19 tweets in R

Suggested Material

- https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1
- https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32
- https://www.mjdenny.com/Text_Processing_In_R.html
- https://www.tidytextmining.com
- https://monkeylearn.com/blog/introduction-to-topic-modeling/
- https://www.youtube.com/watch?v=T05t-SqKArY
- https://www.youtube.com/watch?v=1c_0Ly7tUxY
- https://web.stanford.edu/~jurafsky/slp3/24.pdf
- https://cloud.ibm.com/docs/assistant?topic=assistant-index
- https://www.youtube.com/watch?v=eAncgjQjqIE
- https://hackernoon.com/text-processing-and-sentiment-analysis-of-twitter-data-22ff5e51e14c