Introduction to Natural Language Processing (NLP)

LECTURE 1

Lisa Andreevna Chalaguine

Lecture

About me and my Research

What is NLP

Importance/Applications of NLP

Challenges

Practical

Python Libraries needed

Choice of Dataset

Text Cleaning and Preprocessing

- Stopword removal
- Tokenisation
- Punctuation Removal
- Lemmatisation/Stemming

Lisa A. Chalaguine

PhD Student (Intelligent Systems Group)

Develop Chatbots that argue with people

Converted from Law to Computer Science

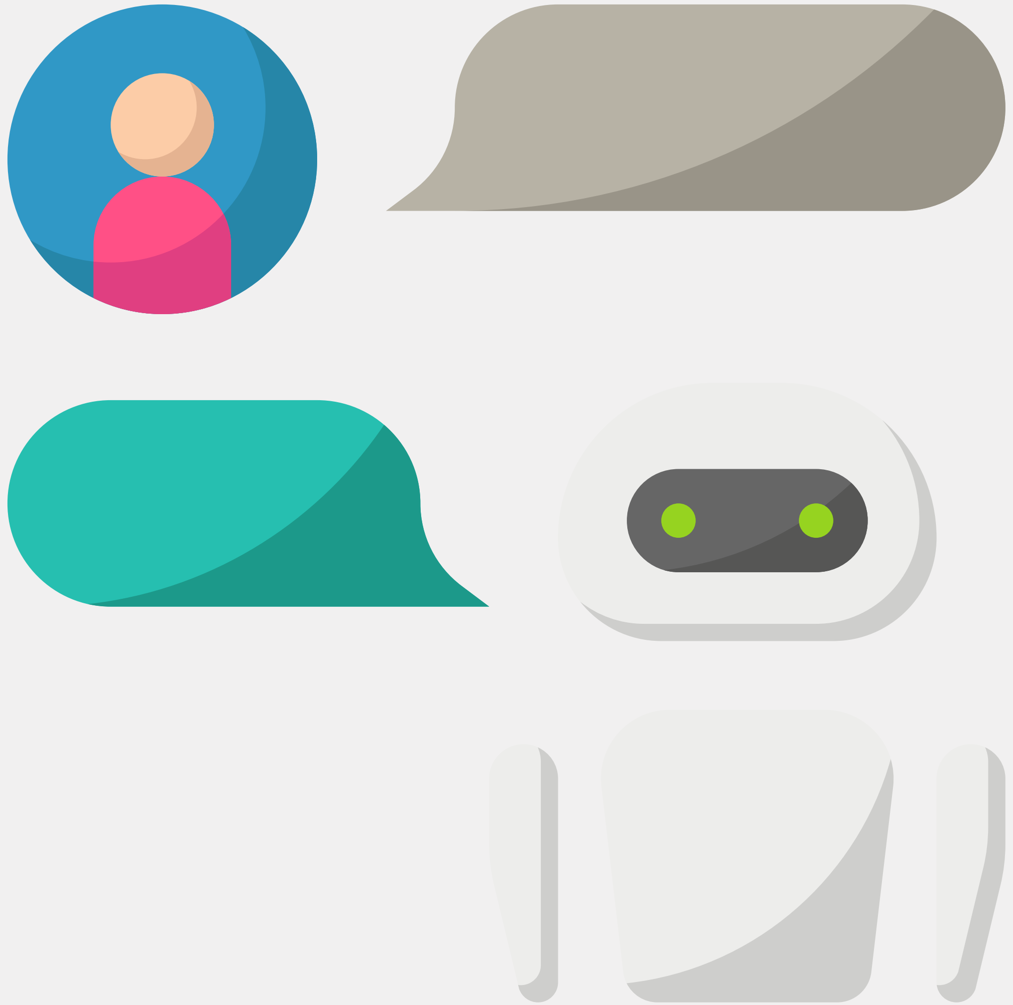
Discovered that being a Penny is better than a being Rachel Zane

Originally from Belarus

And since August most people finally know this country exists...

Demo of latest Project

Chatbot that tries to convince people to get a COVID-19 vaccine, once one is developed and becomes available



What is NLP?

NLP is a frield of AI that gives machines the ability to read, understand and derive meaning from human languages

Discipline that focuses on the interaction between data science and human language.

Programming computers to process and analyse large amounts of natural language data.

Applications

Sentiment/Opinion Analysis (e.g. social media, product reviews...)

Chatbots/Virtual Assistants (Siri, Alexa, Cortana, Google Assistant...)

- Speech Recognition
- Natural Language Generation

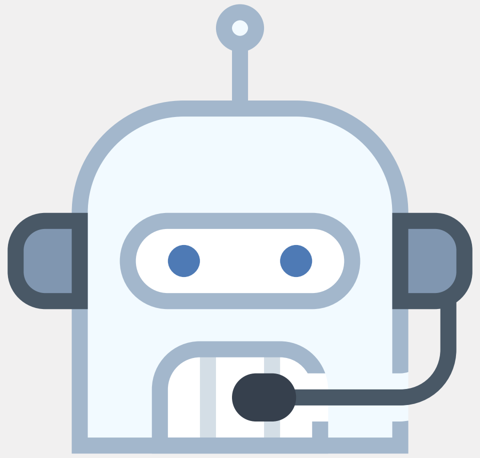
Text Classification (e.g. spam filtering)

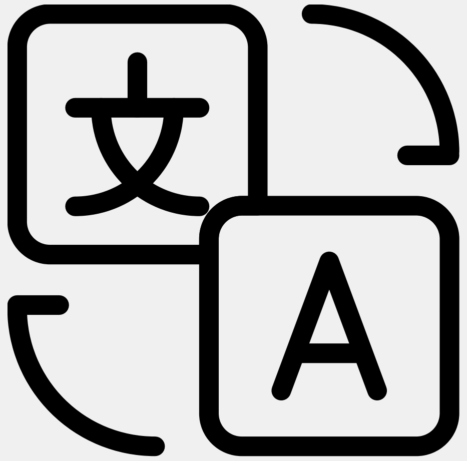
Information extraction

Machine Translation

Text Summarisation

Auto-Correct





Challenges of NLP

Unstructured data

Ambiguity of language (same word - different meaning)

Synonymy (same meaning - different words)

Coreference (what/who do pronouns refer to in subsequent sentences)

Irony, Sarcasm...

Any questions so far?

Practical - stuff you need

Firstly you need ...

- A laptop
- Python 3.6+
- Jupyter Notebooks

Python libraries (for now)

- Pandas
- NLTK

Anaconda comes with most the required libraries and Jupyter!!

Python libraries for next week

- scikit-learn

(requires NumPy & SciPy)

Practical - GitHub page

I will post all of the materials on the following GitHub page

https://github.com/lisanka93/UCL_F2F_NLP101

Try yourself

Select one of the datasets in the dropbox datasets folder and apply some of the preprocessing techniques yourself

For next week

Find a (bigger, more complex) corpus that interests you and that can be used for classification.

NOTE - it will probably not be in .csv format. Your homework is then to write some code to read the corpus into Python