Edge Hill University

</TECHUP_WOMEN>

Natural Language Processing

Introduction

Durham University

Edge Hill University

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

University of York

institute of CODING

Office for Students

Overview

- Natural Language Processing (NLP)
  - What is it?
  - Why?
  - Basic NLP methods

- Collocation Extraction

What is NLP?

- *Ability of computers to understand text*
- *Field of Artificial Intelligence that aims to extract useful information from text written in natural/human language*

Can computers really understand human language?

Children make delicious snacks

Stolen painting found by a tree

Court to try shooting defendant

Ban of nude dancing on Governor's desk

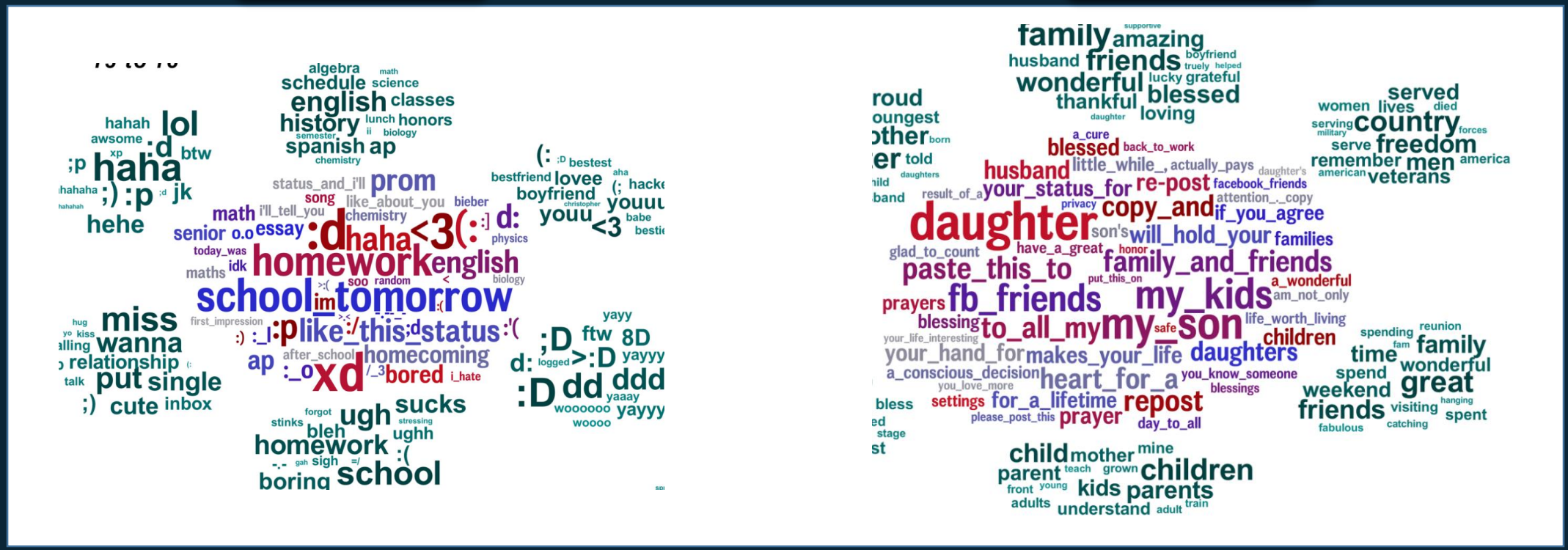
This is awful! (worthy of awe or horrible?)

He has a myriad of cows (10,000 or just a lot?)

Can computers understand text? (Language used in Facebook)

Age 13-18

Age 30-65



Schwartz, H. Andrew, et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach." *PloS one* 8.9 (2013): e73791.

Can computers understand text? (Language used in Facebook)

Extroverts

Introverts

Schwartz, H. Andrew, et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach." *PloS one* 8.9 (2013): e73791.

Can computers really understand human language?

- NLP is hard
  - Natural Language is ambiguous, dynamic and context specific
  - Text can mean different things to different people

- Machine learning to make text mining more efficient
  - Context aware (disambiguation)
  - Generalising observations to new documents
  - Data-driven

Why do we use NLP for Data Analysis?

- Most information is available in natural language (e.g., facebook posts, tweets, blogs, customer reviews, web pages etc.) and not in structured databases
- We cannot query natural language to extract information

How to extract number of negative tweets?

This is just so cute and a great app for little ones.
I hate this app.
This app is soooo slow!
Oh, how my little grandson loves this app.

Tweets relevant to my app (Unstructured Information)

Select count(*) from tweets where tweet='negative'

Why do we use NLP for Data Analysis?

Solution: Use NLP to first convert unstructured to structured data

NLP

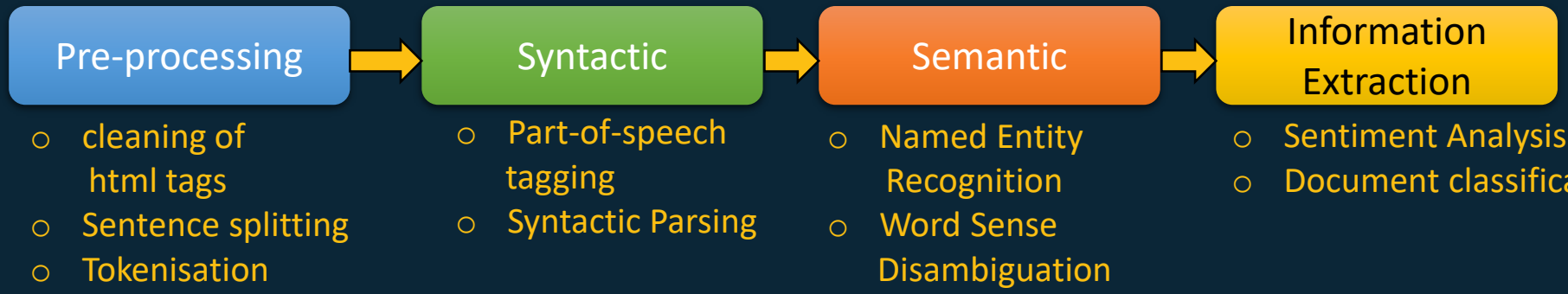
This is just so cute and a great app for little ones.	Positive
I hate this app.	Negative
This app is soooo slow!	Negative
Oh, how my little grandson loves this app.	Positive

Select count(*) from tweets where tweet='negative'

How NLP method are organised?

- NLP applications are organised into hierarchical structures where components of the lower levels inform components of the higher levels (chained applications/worklfows)

- NLP components can be classified into four categories:



Sentence Splitting

- Sentence splitting is a basic pre-processing step that identifies the boundary of sentences

Pre-processing

Easy problem

Tokenisation

- Tokenisation is a pre-processing step that identifies the boundaries of **words**

Pre-processing

Easy problem (e.g., European languages)

Difficult problem for Chinese/Japanese

He is the first African American to have served as president. He previously served in the U.S. Senate representing Illinois from 2005 to 2008 and in the Illinois State Senate from 1997 to 2004.

→ Tokenisation →

He | is | the | first | African | American | to | have | served | as | president. | He | previously | served | in | the | U.S. | Senate | representing | Illinois | from | 2005 | to | 2008 | and | in | the | Illinois | State | Senate | from | 1997 | to | 2004.

他是第一位担任总统的非裔美国人。**他曾于2005年至2008年在美国参**议院代表伊利诺伊州**1997年至2004年在伊利**诺伊州参议院任职

Some languages don't use special character delimiters (e.g., white space) to separate words

Part-of-speech Tagging (POS)

- POS tagging determines the part-of-speech of a word
- Used in many semantic and information-level NLP applications

Syntactic

But mostly solved using machine learning

He is the first African American to have served as president.

→ POS Tagger →

He/**PRP** is/**VBZ** the/**DT** first/**JJ** African/**NNP** American/**NNP** to/**TO** have/**VB** served/**VBN** as/**IN** president/**NN** ./.

JJ
VB VB MD VB
Visiting relatives can be boring.

POS tagging is difficult due to ambiguity

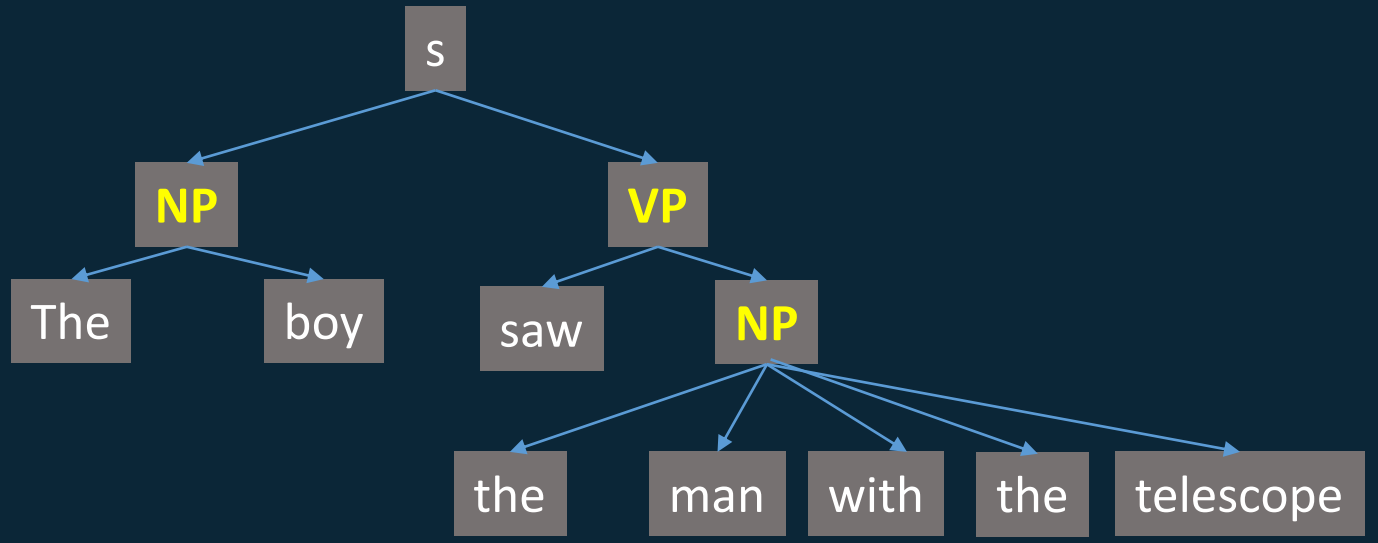
Syntactic Parsing

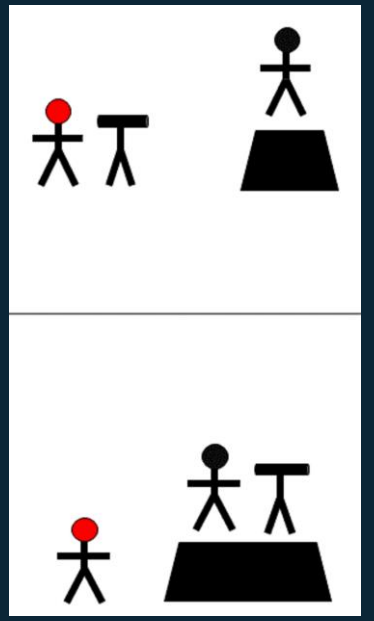
- Identify syntactic structure of a sentence
- Used in word-processors for grammar checking, in information extraction, question answering, word sentence disambiguation and many more!

Syntactic

Good results but still an open problem

The boy saw the a man with a telescope





Sentiment Analysis

Sentiment analysis methods aim to identify and quantify the sentiment expressed in textual content

Also referred to as: Opinion extraction, Opinion Mining, Sentiment Mining, Subjectivity analysis

- Movie: is this movie positive or negative?
- Products: what do people think about the new iPhone?
- Politics: what do people think about a candidate?

Prediction: predict election outcomes or market trends

Sentiment Analysis

This is just so cute and a great app for little ones.	Positive
I hate this app.	Negative
This app is soooo slow!	Negative
Oh, how my little grandson loves this app.	Positive

Difficult problem

Sentiment Analysis

What about sarcasm?

Oh yeah, this movie was TOTALLY great!!. :rolleyes:

Activity 1

- Complete Activity 1
  - Follow the instructions on activity handout