Press-Release Classification && Sorting

Tatiana Fomicheva
MSc in Big Data

Wood Mackenzie

UNIVERSITY of STIRLING

1 Scraping

Identification of the common link part to the documents;
PDF/PPT/Word automated download;
Conversion to .TXT
Tools: BeautifulSoup, Mechanize

2 Tokenization

Separate the PR's in words/per document
Tools: NLTK

3 Pre-Processing

Remove specific common news identifiers and custom stop-words;
Stemming and Lemmatization
Tools: NLTK

4 Feature Engineering

Bag of Word OR
Word Embedding OR
Combination of BoW & Word Embedding
Tools: TF-IDF, Word2Vec, GloVe, FastText

5 Model (Classifier) Building

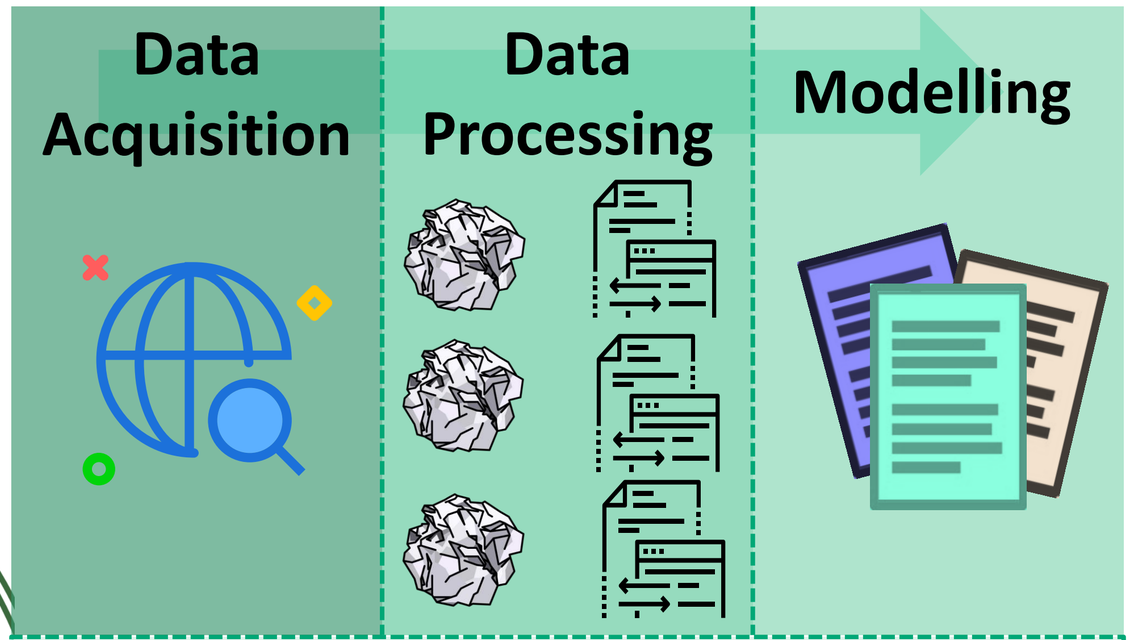
Supervised Models testing
Tools: Pipeline, RandomForest, GradientBoosting, ExtraTrees, Bagging, Kneighbours, MLP, LogReg, ...

6 Vectorizer&&Model Tuning

Hyperparameters Tuning;
Weight adjustment, reducing false negatives
Tools: GridSearchCV, Pipeline, Matplotlib, ...

7 Model Evaluation&&Extension

Accuracy testing;
Generalize to be used for Press Releases from different companies
Tools: Sklearn metrics



Goal

Create a generalized framework which automates the text acquisition from different sources, pre-processing and modelling to attribute labels per text file accurately

For further information contact: *Tatiana Fomicheva* *taf00014@students.stir.ac.uk*