TEXT MINING

BY ANDREA CARDINALI & ADONIS KINGSLEY GRANITA

GOALS:

This project aims to perform two text minings tasks: text classification and text summarisation.

Two different datasets are used for the tasks.

TEXT SUMMARIZATION

DATASET

SAMPLE of 20000 stories
from the CNN dataset

stories contained:

- actual story
- summmmary sentences

story	highlights
OTTAWA (CNN) -- Canadian officials say they ha...	[Canadian officials say they were following up...
(CNN) -- Rick Perry's decision to pull out of ...	[Paul Sracic: If Mitt Romney wins in S.C., the...
Comedian Joan Rivers isn't about to apologize ...	[Rivers made the quip Monday on a show about O...
(CNN) -- A Southern California man who attract...	[Sloan Steven Briles, 35, of Irvine, Californi...
(CNN) -- Verizon pulled a rabbit out of its co...	[Writers: Verizon, AT&T deals are attempts to ...
...	...
Baghdad (CNN) -- Wearing hooded sweatshirts, b...	[Smashing Hits is an English-language rap grou...
(CNN) -- At least 150 people drowned when a bo...	[Thousands of refugees from fighting in Libya ...
(CNN) -- Athletic Bilbao striker Fernando Llor...	[Juventus open talks with Athletic Bilbao's Fe...
(CNN)"Let us build a fairyland for the people ...	["Let the strong wind of fish farming blow acr...
(CNN) -- Australia has a new prime minister. A...	[Australian PM's Labor Party voted her out of ...

PREPROCESSING

PREPROCESSING

- stop_words removal
- punctuation
- lemmatization
- normalization of text

TEXT REPRESENTATION

- TF-IDF

EVALUATION

ROUGE-1

- Measures the overlap of unigrams (single words)

ROUGE-2

- Measures the overlap of bigrams (contextual accuracy)

ROUGE-L

- Measures the Longest Common Subsequence (LCS)(words that appear in the same order)

LSA

SVD

- U (left singular vectors):Represents sentences projected in a topic space.
- Σ (diagonal matrix): Contains singular values that indicate the importance of each latent topic.
- V^T (right singular vectors): Represents the terms' relationships with the latent topics.

LSA

SCORE

$$\text{score}_i = \sum_{j=1}^{n} (\text{LSA}_i[j])^2$$

$$LSA = U \cdot \Sigma$$

TEXT RANK

COSINE SIMILARITY

- A cosine similarity matrix is computed from the TF-IDF matrix.
- To capture the pairwise relationships (similarity) between sentences.

TEXT RANK

GRAPH REPRESENTATION

- The similarity matrix is converted into a graph using the NetworkX library:
- Sentences are represented as nodes.
- Cosine similarity values between sentences are represented as edges (weights of the graph).

TEXT RANK

PageRank

- Each sentence (node) is scored based on its connectivity and the importance of the sentences it is connected to.
- This algorithm assigns higher scores to sentences that are connected to other highly-ranked sentences.

SCORING

SUMMARY

1. Sentences are ranked by their scores in descending order.
2. The top 3 sentences with the highest scores are selected for the summary

EVALUATIONS

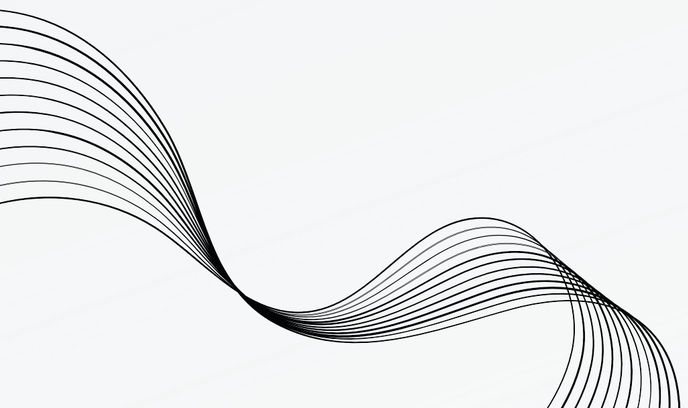
LSA

- ROUGE-1: 0.21
- ROUGE-2: 0.05
- ROUGE-L: 0.17

TEXT RANK

- ROUGE-1: 0.26
- ROUGE-2: 0.08
- ROUGE-L: 0.18

TEXT CLASSIFICATION



THE DATA

This collection contains more than 190K texts about legit and spam emails. The original data is available here.

	label	text
0	Spam	viiiiiiagraaaa\nonly for the ones that want to...
1	Ham	got ice thought look az original message ice o...
2	Spam	yo ur wom an ne eds an escapenumber in ch ma n...
3	Spam	start increasing your odds of success & live s...
4	Ham	author jra date escapenumber escapenumber esca...
...	...	...
193847	Ham	on escapenumber escapenumber escapenumber rob ...
193848	Spam	we have everything you need escapelong cialesc...
193849	Ham	hi quick question say i have a date variable i...
193850	Spam	thank you for your loan request which we recie...
193851	Ham	this is an automatically generated delivery st...

PRE-PROCESSING

To apply text classification we had to first pre-process all the texts:

The target variable was re-encoded to be ‘spam’ = 1 and ‘ham’ = 0.

It was then chosen a sample of 20K emails to lower the resources needed for the models.

Each email’s text was vectorized using a count vectorizer and later a tf-idf vectorizer.

	label	text	target
0	Spam	viiiiiiagraaaa\nonly for the ones that want to...	1
1	Ham	got ice thought look az original message ice o...	0
2	Spam	yo ur wom an ne eds an escapenumber in ch ma n...	1
3	Spam	start increasing your odds of success & live s...	1
4	Ham	author jra date escapenumber escapenumber esca...	0
...	...	...	...
193847	Ham	on escapenumber escapenumber escapenumber rob ...	0
193848	Spam	we have everything you need escapelong cialesc...	1
193849	Ham	hi quick question say i have a date variable i...	0
193850	Spam	thank you for your loan request which we recie...	1
193851	Ham	this is an automatically generated delivery st...	0

193852 rows × 3 columns

APPLYING THE CLASSIFICATION

Decision Tree

Based on a tree-like structure. Starting from a root node, it splits data into subsets.

SVC

It tries to find the best hyperplane (maximises the distance) that separates the classes in the feature space.

Random Forest

This is based on a collection of decision trees. It combines the outputs of multiple decision trees for the prediction.

EVALUATION

ACCURACY

MEASURES THE MODEL ABILITY TO CORRECTLY PREDICT THE CLASS

RECALL

MEASURES THE PORTION OF ACTUAL POSITIVE INSTANCES CORRECTLY IDENTIFIED

F1

WEIGHTED AVERAGE OF ACCURACY AND RECALL

APPLYING THE CLASSIFICATION

Decision Tree

Accuracy: 0.9102
F1 score: 0.903
Recall score: 0.902

SVC

Accuracy: 0.97
F1 score: 0.968
Recall score: 0.978

Random Forest

Accuracy: 0.9566
F1 score: 0.952
Recall score: 0.937

STRATIFIED 5-FOLD CLASSIFICATION

This technique is used to ensure that the evaluation is not biased. The data is divided into subsets and the performance metrics are averaged.

It was applied to the data using the same three models as before. After five iterations the F1 measures were averaged to get the single final measure.

5-FOLD CLASSIFICATION

The results were very satisfying:

```
Stratified K-Fold with 5 splits gives the following
Decision tree: 0.90 +- 0.005
SVM tree:      0.97 +- 0.003
RF tree:       0.95 +- 0.002
```

These are very much similar of the single models. The subsets do not differ much from the full dataset.

THANKS FOR YOUR ATTENTION