Giannelli Alessio
Imbonati Lorenzo
Valoti Davide

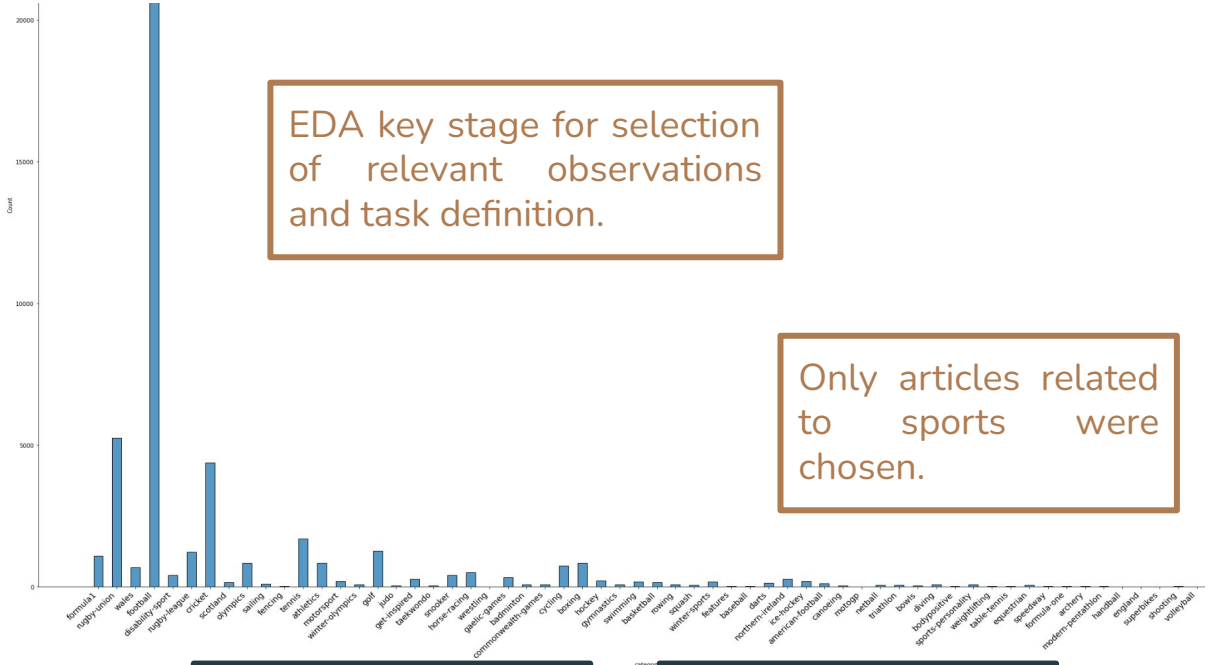
X-SUM(MARIZATION)

TEXT ANALYSIS OF BBC NEWS ARTICLES

Data Science Master's Degree
Text Mining Exam

INDEX

EXPLORATORY DATA ANALYSIS



EDA key stage for selection of relevant observations and task definition.

Only articles related to sports were chosen.

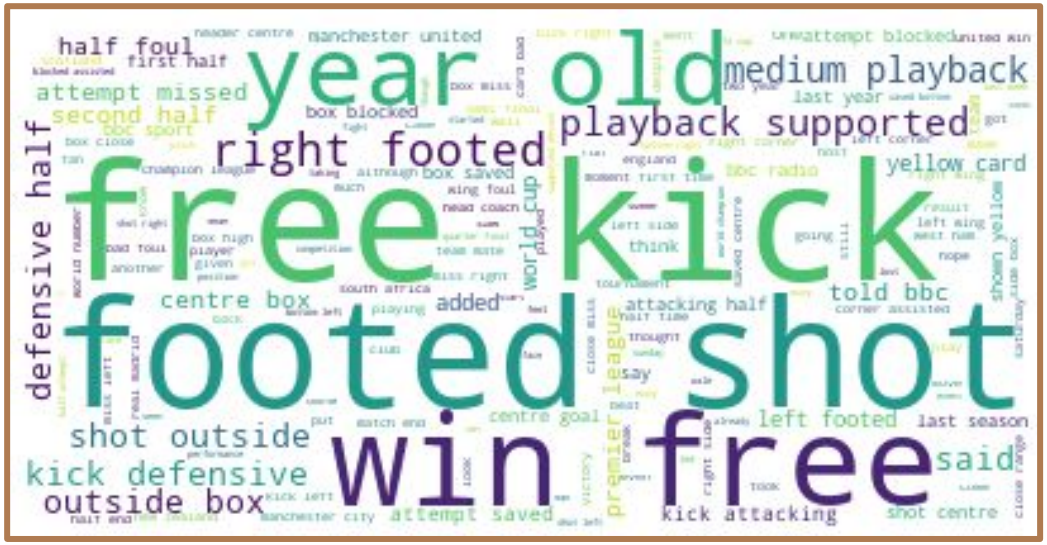
Original dataset:
226,711 documents

Filtered dataset:
55,416 documents

CATEGORY	DOCUMENT
Football	25718
Rugby Union	5253
Cricket	4370
Tennis	1693
Golf	1255
Rugby League	1227
Formula 1	1081
Boxing	844
Olympics	834
Athletics	831
Cycling	741
Wales	681
Horse racing	513

EXPLORATORY DATA ANALYSIS

Wordcloud



WORD	FREQUENCY
Free kick	1.00
Footed shot	0.81
Win free	0.80
Year old	0.57
Right footed	0.56

PRE-PROCESSING

Text Homogeneity

White space removal
Spelling correction
Emoji, links and HTML tags removal

Lower case conversion

```
qundogan, 26, told bbc sport he "can see the finis
llowing back surgery that kept him out for a year,
start of the premier league season at brighton on
to fall and fight your way back. you feel good and
```

Symbol and punctuations removal

```
gundogan  26  told bbc sport he  can see the finis
llowing back surgery that kept him out for a year
start of the premier league season at brighton on
to fall and fight your way back  you feel good and
```

Numbers removal

```
 gundogan    told bbc sport he  can see the finishi
ack surgery that kept him out for a year  and sat c
e premier league season at brighton on  august but
ght your way back  you feel good and feel ready  th
```

PRE-PROCESSING

Tokenization

[ferrari, appeared, in, a, position, to, challenge, until, the, final, laps, when, the, mercedes, stretched, their, legs, to, go, half, a, second, clear, of, the, red, cars, sebastian, vettel, will, start, third, ahead, of, team, mate, kimi, raikkonen, the, world, champion, subsequently, escaped, punishment, for, reversing, in, the, pit, lane, which, could, have, seen, him,

Lemmatization

[ferrari, appeared, in, a, position, to, challenge, until, the, final, lap, when, the, mercedes, stretched, their, leg, to, go, half, a, second, clear, of, the, red, car, sebastian, vettel, will, start, third, ahead, of, team, mate, kimi, raikkonen, the, world, champion, subsequently, escaped, punishment, for, reversing, in, the, pit, lane, which, could, have, seen, him,

Bigrams and trigrams

[ferrari, appeared, in, position, to, challenge, until, the, final, lap, when, the, mercedes, stretched, their, leg, to, go, half, second, clear, of, the, red, car, sebastian_vettel, will, start, third, ahead, of, team_mate, kimi_raikkonen, the, world_champion, subsequently, escaped_punishment, for, reversing, in, the, pit_lane, which, could, have, seen, him,

Stop words removal

[ferrari, appeared, position, challenge, final, lap, mercedes, stretched, leg, go, half, second, clear, red, car, sebastian_vettel, start, third, ahead, team_mate, kimi_raikkonen, world_champion, subsequently, escaped_punishment, reversing, pit_lane, could,

TOPIC MODELING
LDA - Input

Dictionary & Filter

```
1 dictionary = gensim.corpora.Dictionary(df['doc_trigram_lem'])
2
3 #TO BE INCLUDED:
4 #no_below: minimum number of documents that a token must appear in
5 #no_above: maximum fraction of documents that a token can appear in
6 dictionary.filter_extremes(no_below=5, no_above=0.5, keep_n=100000)
```

```
(49361 unique tokens: ['able', 'action', 'adding', 'admitting', 'agreement']...)
```

Doc Representation

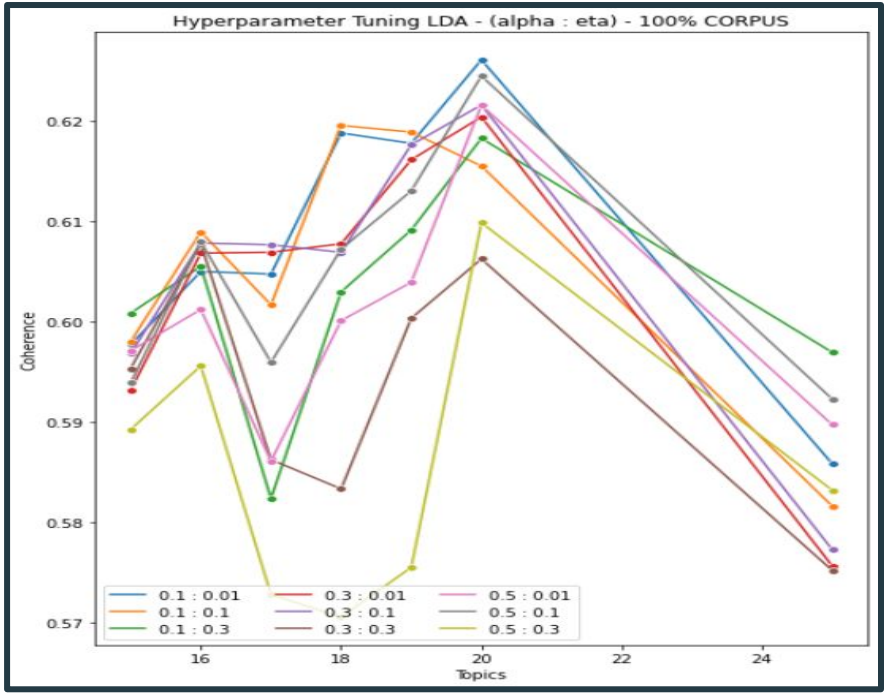
Bag of Words

```
[('able', 1),
 ('action', 1),
 ('adding', 1),
 ('admitting', 1),
 ('ago', 1),
 ('agreement', 1),
 ('ahead', 3),
 ('alonso', 1),
 ('already', 1),
 ('also', 1),
 ('appeared', 2),
 ('arrived', 1),
 ('attempt', 1),
 ('australia', 2),
 ('back', 1),
```

TOPIC MODELING
LDA - Hyper-parameters tuning

The tuning phase was driven by CV_coherence (y-axis) and the number of the latent topics (x-axis).

In this phase we have optimized alpha and beta, the hyperparameters used for the Dirichlet prior distribution that define the document-topic and word-topic relationships.



TOPIC MODELING
LDA - Final model results

```
lda_model = gensim.models.LdaModel(bow_corpus, id2word=dictionary, num_topics=20, offset=2,
random_state=100, update_every=1, passes=10, alpha='0.1', eta='0.01', per_word_topics=True)
```

Coherence CV
0.626

The Coherence CV value of 0.62 indicates that the model has good internal coherence.

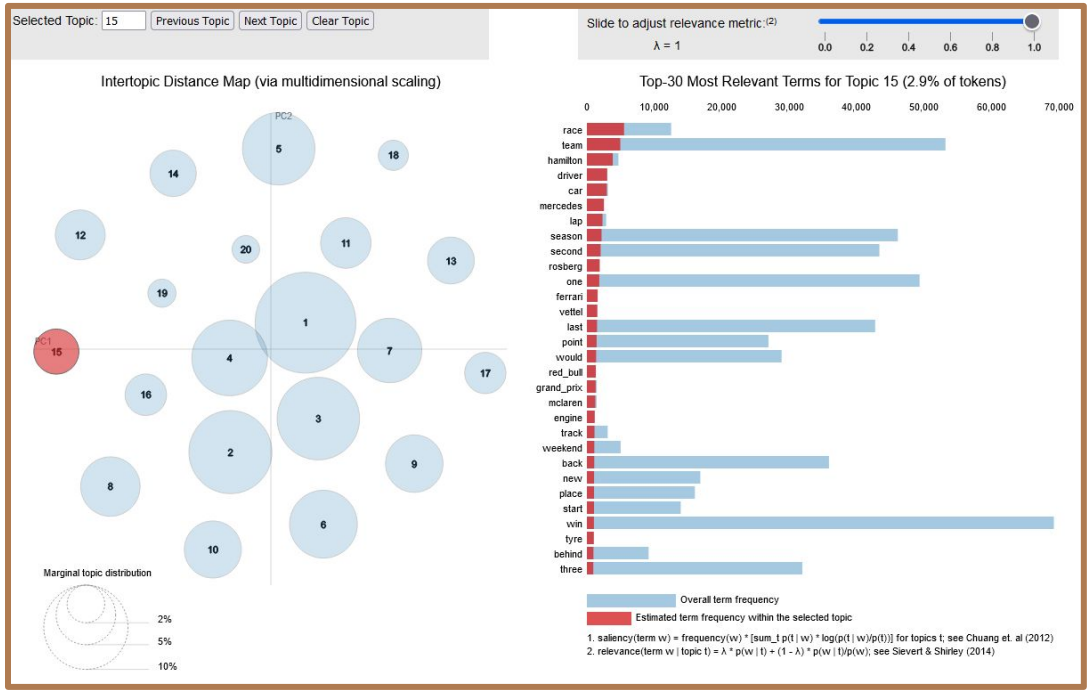
Coherence Umass
-1.723

The Coherence Umass value of -1.72 suggests that the model may have issues with external coherence.

Perplexity
-7.626

The Perplexity of -7.62 indicates that the model has good generalization ability.

TOPIC MODELING
LDA - Evaluation Results



LET'S LEAVE ROOM FOR
LDA_viz.html

TOPIC MODELING - LDA Evaluation Result

Human Evaluation

TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5
/	Football	Foundation of Football	Foundation of Football	Fifa & Decision
TOPIC 6	**TOPIC 7**	**TOPIC 8**	**TOPIC 9**	**TOPIC 10**
Barclays Premier League	Football Tournament	Olympiad	Transfer Market	Rugby
TOPIC 11	**TOPIC 12**	**TOPIC 13**	**TOPIC 14**	**TOPIC 15**
Foundation of Football	Baseball	/	Cricket	Formula1
TOPIC 16	**TOPIC 17**	**TOPIC 18**	**TOPIC 19**	**TOPIC 20**
/	Scottish Premier League	Football League One	African Cup	Women's Super League

Words Distribution

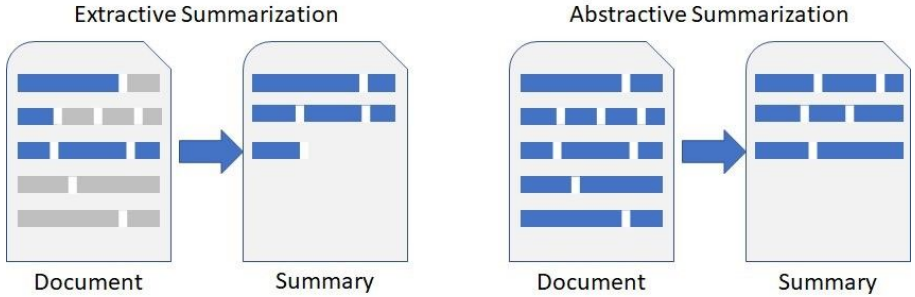
```
'0.021*"race" + 0.019*"team" + 0.014*"hamilton" + 0.011*"driver" + '
'0.011*"car" + 0.009*"mercedes" + 0.009*"lap" + 0.008*"season" + '
'0.008*"second" + 0.007*"rosberg" + 0.007*"one" + 0.006*"ferrari" + '
'0.006*"vettel" + 0.006*"last" + 0.005*"point" + 0.005*"would" + '
'0.005*"red_bull" + 0.005*"grand_prix" + 0.005*"mclaren" + 0.004*"engine"'
```

This string represent the 20 top words in Topic 15, sorted by their semantic relevance to the topic.

Each word is accompanied by its weight, represented as a float, which indicates how strongly the word is associated with the topic.

TEXT SUMMARIZATION

Implementation of different **multi-document summarization algorithms,** in order to find the best one for the considered dataset.



Oliver Kebble signs two-year contract with Glasgow Warriors

15 February 2017 | Rugby Union

Kebble has not represented South Africa at senior level

Glasgow Warriors have confirmed the signing of loose-head prop Oliver Kebble for next season.

The 24-year-old has agreed a two-year-deal and will arrive following his commitments with the Stormers and Western Province in South Africa.

He is the son of former Springbok Guy Kebble and won the 2012 Under-20s World Championship with South Africa.

"I try and bring an edge to the game and make an impact," Kebble told the Pro12 club's website.

"I've watched all of the Glasgow matches in the Champions Cup this season and northern hemisphere rugby is getting very exciting. I'm looking forward to playing in a competitive European league.

"I know Dave Rennie is one of the best coaches in the world, so it's an exciting prospect to work under him next season."

Kebble will join current team-mate **Huw Jones** in Glasgow, with the Scotland centre signing a two-year contract with the Warriors earlier this month.

"Huw and I live together in Cape Town," he explained. "We didn't really talk about it too much before it happened, but now it's nice to know there will be a familiar face in Glasgow."

Last week, BBC Scotland revealed Kebble's expected arrival, with the new recruit **considered a project player** by Scottish Rugby, who have monitored him for several years.

TEXT SUMMARIZATION
Extractive approach

- Selection of important phrases or sentences from the input document.
- Naive approach as **benchmark:** selection of the first three sentences of each document.

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
BASELINE	0.168	0.020	0.107	0.107

"The 24-year-old has agreed a two-year-deal and will arrive following his commitments with the Stormers and Western Province in South Africa. He is the son of former Springbok Guy Kebble and won the 2012 Under-20s World Championship with South Africa. "I try and bring an edge to the game and make an impact," Kebble told the Pro12 club's website."

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
BASELINE	0.075	0.000	0.075	0.075

TEXT SUMMARIZATION
Abstractive approach

Architectures used:

- **T5** (Exploring the limits of Transfer Learning with a Unified Text-To-Text Transfer Transformer)
- **BART** (Bidirectional Auto-Regressive Transformers)
- **PEGASUS** (Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence)

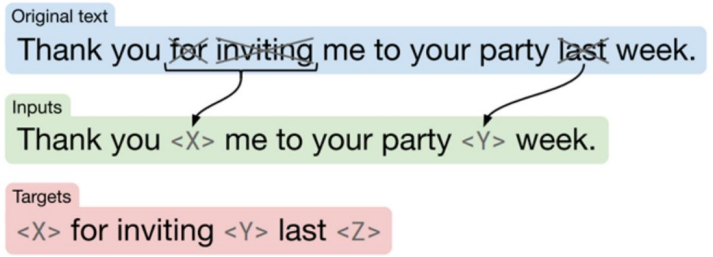


Figure 6: Image taken from "Raffel, Shazeer and Roberts, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"

ABSTRACTIVE SUMMARIZATION
Preprocessing

Use of **AutoTokenizer** for the preprocessing phase.

Main steps:

- Text normalization
- Tokenization
- Byte-pair encoding (BPE)
- Special tokens
- Padding and truncation
- Encoding

```
"Gregor Townsend gave a debut to powerhouse fijian-born wallaby wing Taqele Naiyaravoro . the dragons gave first starts of the season to wing aled Brew and hooker Elliot Dee . it took 24 minutes for a disjointed game to produce a try ."
```

```
[32939, 1422, 114, 4042, 112, 19077, 19716, 46827, 121, 7623, 1075, 304, 1846, 6959, 781, 19854, 12078, 6856, 9162, 32708, 26977, 110, 107, 109, 24674, 1422, 211, 2171, 113, 109, 578, 112, 6959, 114, 4105, 20663, 111, 5922, 420, 29592, 15960, 110, 107, 126, 635, 1202, 542, 118, 114, 62478, 389, 112, 1449, 114, 508, 110, 107, 1]
```

ABSTRACTIVE SUMMARIZATION
Results

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
BASELINE	0.168	0.020	0.107	0.107
T5	0.171	0.023	0.117	0.117
BART	0.203	0.041	0.135	0.166
PEGASUS	**0.472**	**0.269**	**0.412**	**0.414**

PEGASUS obtains best performance for the considered dataset.

ABSTRACTIVE SUMMARIZATION
Single document results

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
T5	0.131	0.000	0.098	0.098

"the 24-year-old has agreed a two-year deal with the Stormers and Western Province . he is the son of former Springbok Guy Kebble and won the 2012 under-20s world championship with south africa . Kebble will join current team-mate Huw Jones in Glasgow ."

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
BART	0.083	0.000	0.083	0.083

"Kebble is the son of former Springbok Guy Kebble and won the 2012 Under-20s World Championship with South Africa. The 24-year-old has agreed a two-year deal with the Pro12 club. He will arrive following his commitments with the Stormers and Western Province in South Africa and will join current team-mate Huw Jones."

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
PEGASUS	**0.320**	**0.174**	**0.320**	**0.320**

"Glasgow Warriors have signed South African scrum-half Ruan Kebble."

ABSTRACTIVE SUMMARIZATION
PEGASUS Fine-tuning

Training set:
10000 documents

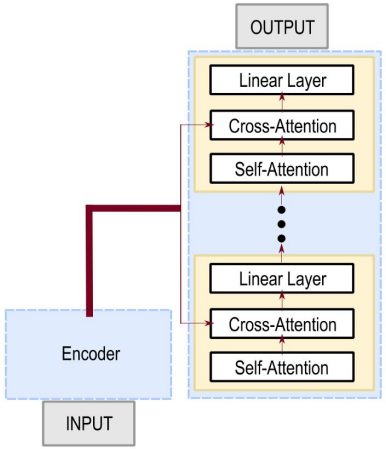
Validation set:
1000 documents

Application of a **Seq2Seq trainer** to estimate parameters adaptive to the new training set (supervised fine-tuning).

Main parameters:

- 8 epochs of training
- Training batch size of 8 documents
- Validation batch size of 4 documents

Encoder-Decoder



ABSTRACTIVE SUMMARIZATION
PEGASUS Fine-tuning results

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
PEGASUS	0.472	0.269	0.412	0.414
PEGASUS FINE-TUNING	**0.497**	**0.275**	**0.418**	**0.418**

Proposal summary

"Glasgow Warriors have signed South African loose-head prop Oli Kebble for next season."

True Summary

"Glasgow Warriors have confirmed the signing of loose-head prop Oliver Kebble for next season"

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L SUM
PEGASUS	0.320	0.174	0.320	0.320
PEGASUS FINE-TUNING	**0.690**	**0.519**	**0.690**	**0.690**

CONCLUSIONS & FUTURE DEVELOPMENTS

Topic Modeling	Through the **LDA** model it was possible to identify the main Topics and the related terms that characterize the various documents.
Summarization	**PEGASUS** model the best to implement Extractive and Abstractive summarization, and **Fine-Tuning** has further improved performance.
Future developments	Results of topic modeling as input for Extractive Text Summarization to make a better comparison between Extractive and Abstractive summarization.