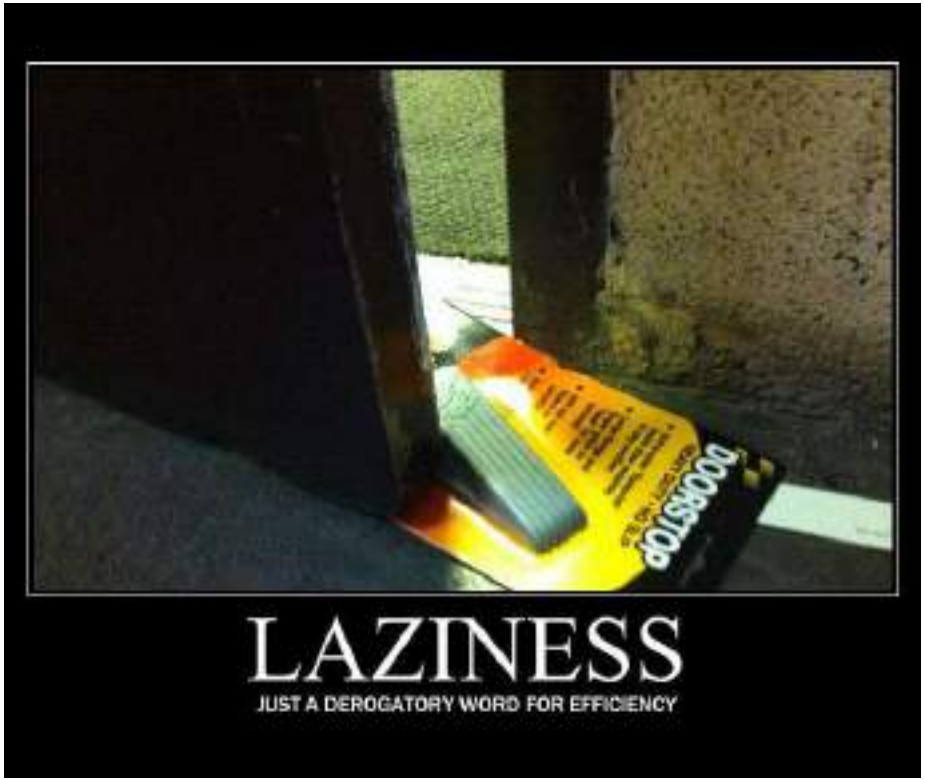

tl;dr
Automatic Text Summarization

I would have written a shorter letter,

but I didn’t have the time.

*—Blaise Pascal, Mathematician*

# Why is this project relevant?




Unprecedented access to content
So little time to read it

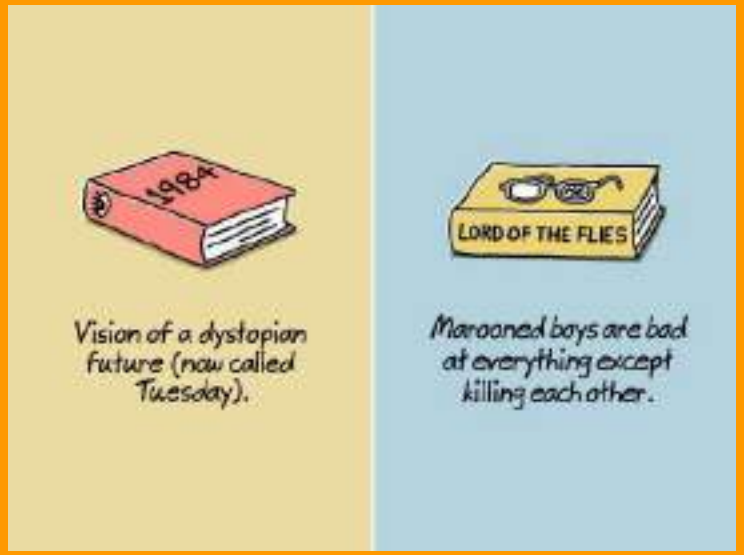
You've prepared a report
Boss only wants the main points

You're a researcher
Can't be bothered with abstract writing

• • •

# Extractive

**Task:** Highlight most important points




# Abstractive

# What makes a good summary?

**Goals:**

- Optimize for topic coverage
- Optimize for readability

**Evaluation criteria:**

- Salience
- Length
- Structure & coherence
- Balance
- Grammar
- Non-redundancy

**Most Popular Metric:**

Rouge-N - Form of Recall of n-grams

**Best Metric:**

Asking users:
“Does this summary answer your question?”

# Setup of the Project

**Training Data:** CNN/Daily Mail Dataset
312.085 News Articles and human written Reference Summaries

**Metric:**
Rouge-1 Score

**Implemented Models:**

1. Extractive Model - Textrank
2. Abstractive Model - Seq2Seq
3. Mixed Model - Pointer Generator

# Extractive Algorithms

## Positional Method

Game of Thrones is an American fantasy drama television series created by David Benioff and D. B. Weiss for HBO. It is an adaptation of A Song of Ice and Fire, George R. R. Martin's series of fantasy novels, the first of which is A Game of Thrones. The show was both produced and filmed in Belfast and elsewhere in the United Kingdom. Filming locations also included Canada, Croatia, Iceland, Malta, Morocco, and Spain.
The series premiered on HBO in the United States on April 17, 2011, and concluded on May 19, 2019, with 73 episodes broadcast over eight seasons.

# Textrank

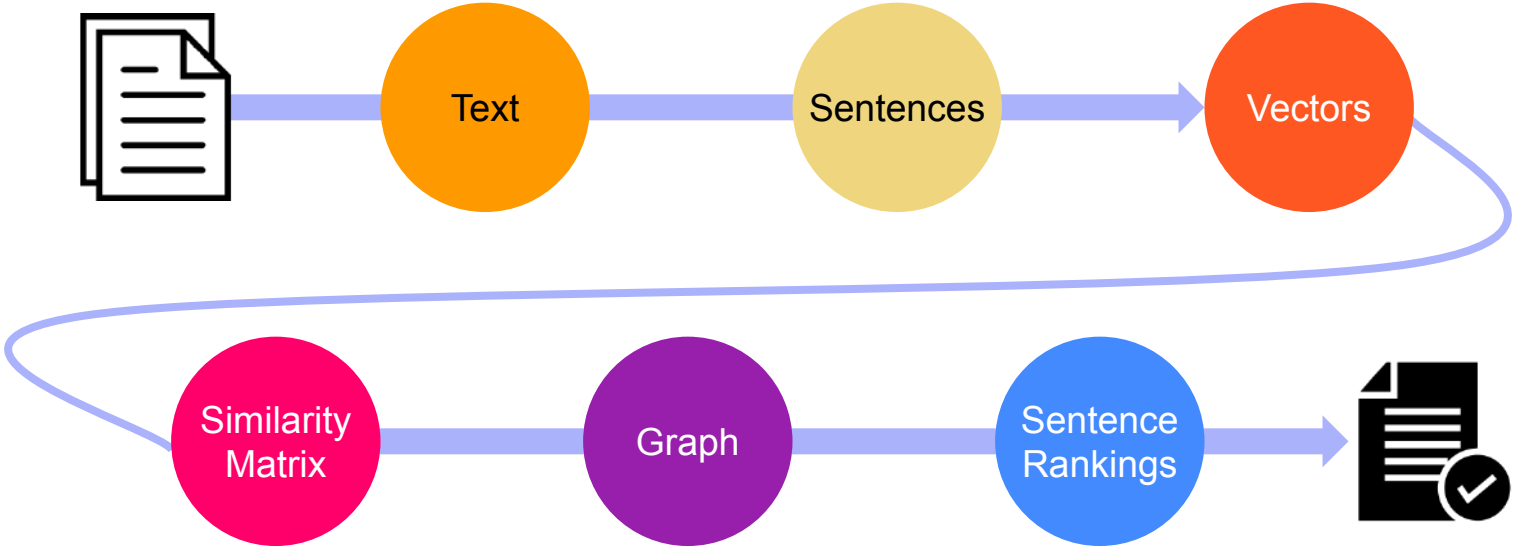
**Authors:**

“...based on the concept of eigenvector centrality in a graph representation of sentences”, using “a connectivity matrix based on on intra-sentence cosine similarity...”

**Plain English:**

- Finds the relative importance of all words in a document
- Selects sentences which contain the most of those high-scoring words

# How it works


Text
Sentences
Vectors
Similarity Matrix
Graph
Sentence Rankings

# TextRank: Test Summaries

**Article Oil Crisis**

**Article Climate**

**Book Non-Fiction**

**Book Fiction**

**Biography**

Riyadh, Saudi Arabia (CNN) Saudi and US investigators have determined "with very high probability" that the weekend attack on the Saudi oil industry was launched from an Iranian base in Iran close to the border with Iraq, according to a source familiar with the investigation.
Another source who has spoken with Saudi government officials has told CNN that based on images of the wreckage that fell in the desert, at least some of the missiles used are known as the Quds 1.

# TextRank: Test Summaries

**Article Oil Crisis**

**Article Climate**

**Book Non-Fiction**

**Book Fiction**

**Biography**

Do not, however, mistake this as a plea to doctors to start prescribing more sleeping pills—quite the opposite, in fact, considering the alarming evidence surrounding the deleterious health consequences of these drugs.
After twelve to eighteen months of no sleep, the patient will die.
Though exceedingly rare, this disorder asserts that a lack of sleep can kill a human being.

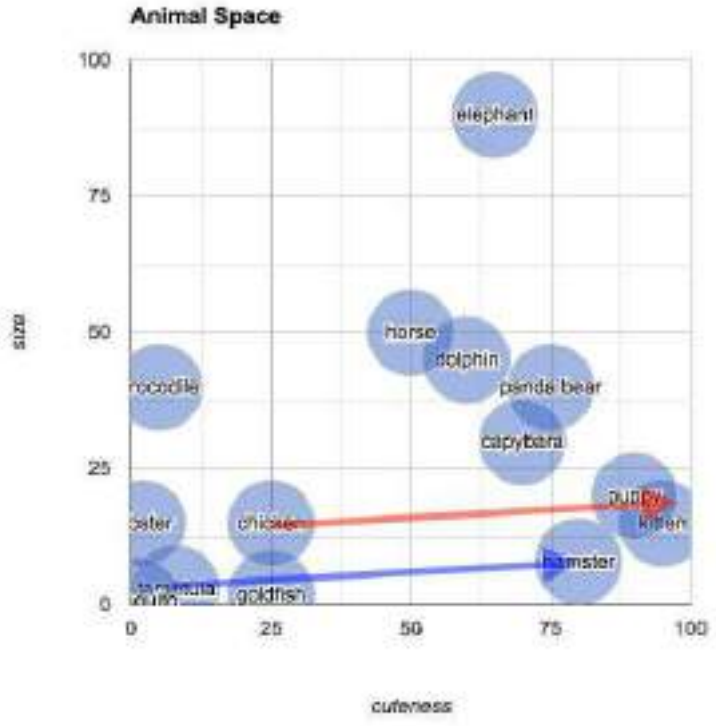
# Abstractive Summarization

Teach Neural Network to generate words, rather than copying from the input text

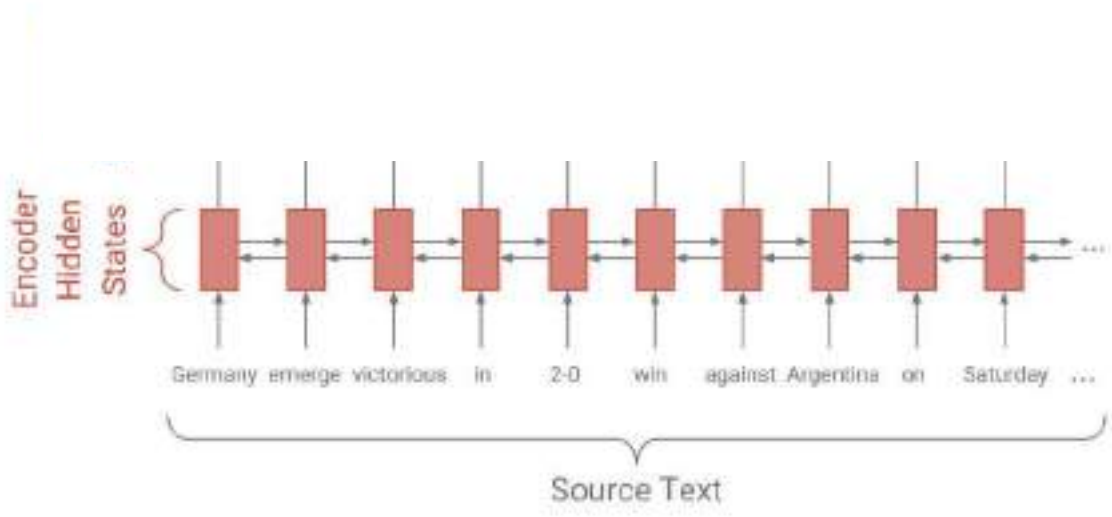
**Input:**
Turn Words into Vectors (Word-Embedding)

Man - Woman + Queen = King

Munich - Dortmund + Best + Football = ?



# Abstractive - Seq2Seq



# Seq2Seq: Test Summaries

**Article Oil Crisis**

missile attack on <UNK> is
said to have been said to be

**Article Climate**

**Book Non-Fiction**

a new <UNK>
a <UNK> sleep
diet hormone is a <UNK>
sleep is dying
doctors to start <UNK> <UNK>
<UNK> <UNK>
cause of fatal accident
in all developed nations
doctors to start medical advice
genetic disorder

**Book Fiction**

**Biography**

# Problem Patterns

Expensive to train and does the Neural Network really understand?

**Network can’t copy facts**

as it doesn’t copy words, but generates them, it sometimes is incapable of generating facts correctly

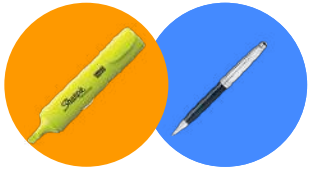
“sleep is dying”

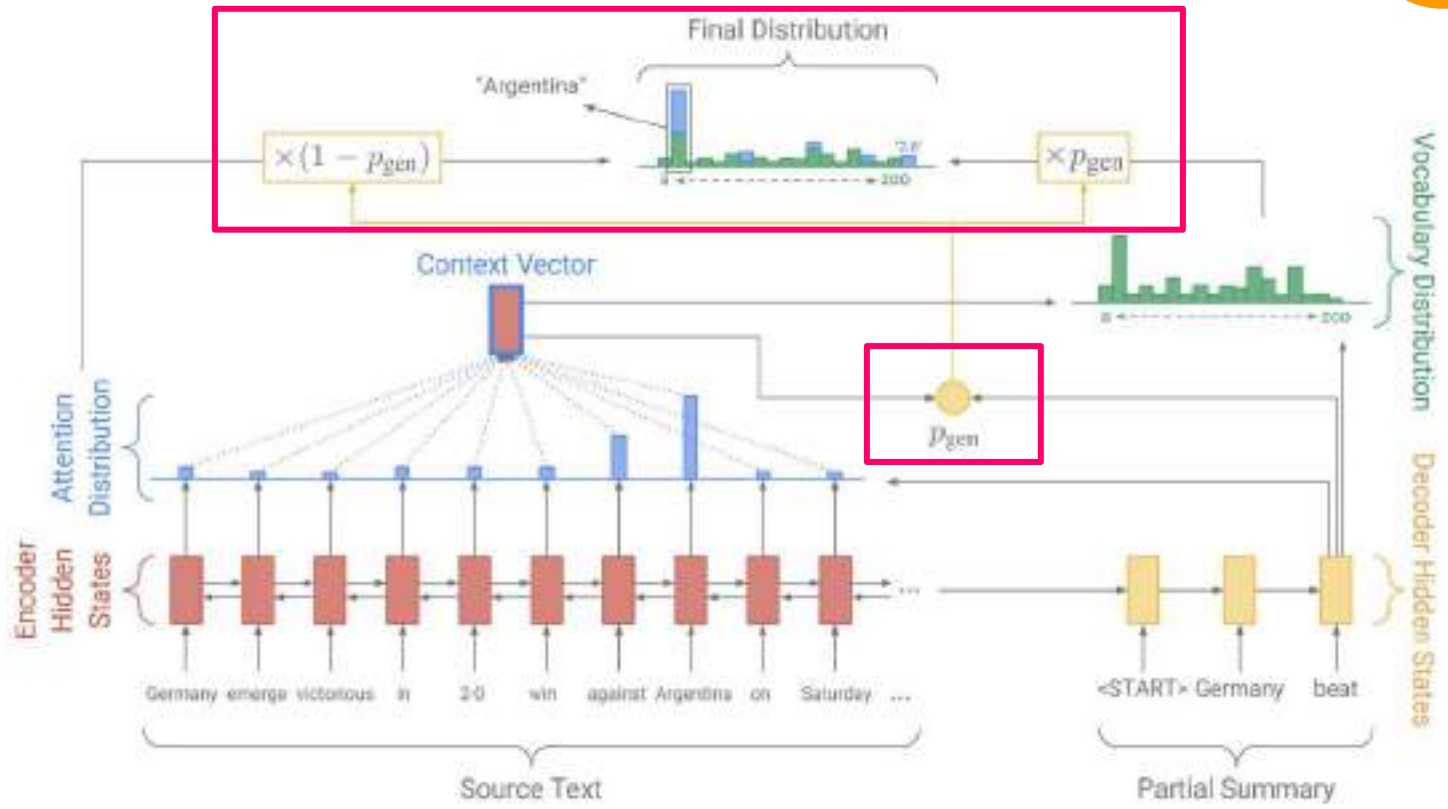
**Repeats words**

Could be that the network relies too much on decoder input (e.g. summary word from before)

“scientists say
scientists say
scientists say
climate change”

# Mixed - Pointer Generator





# Poin-Gen: Test Summaries

- Article Oil Crisis
- Article Climate
- Book Non-Fiction
- Book Fiction
- Biography

two-thirds of adults throughout all developed nations fail to obtain the recommended eight hours of nightly sleep.i . insufficient sleep is a key lifestyle factor determining whether or not you will develop alzheimer 's disease . inadequate sleep is a key lifestyle factor determining whether or not you will develop alzheimer 's disease .

# Poin-Gen: Test Summaries

- Article Oil Crisis
- Article Climate
- Book Non-Fiction
- **Book Fiction**
- Biography

de sade was one of the most gifted and abominable personages in an era that knew no lack of gifted and abominable personages . his name was jean-baptiste grenouille , and if his name , in contrast to the names of other gifted abominations , de sade 's , or saint-just 's , bonaparte 's , bonaparte 's , has been forgotten today , it is certainly not because grenouille fell short of those more famous blackguards .

| | Textrank | Seq2Seq | Pointer Generator |
|---|---|---|---|
| Article Oil Crisis | 👍 | 👎 | 👍 |
| Article Climate | 👍 | 👎 | 👍 |
| Book Non-Fiction | 👎 | 👎 | 👍 |
| Book Fiction | 👎 | 👎 | 👎 |
| Biography | 👍 | 👎 | 👍 |
| ROUGE 1 - Score | 32.4 % | 6.1 % | 39.1 % |

# What’s next?

**Bottleneck is quality data**
Working on scraping together a data set for book summarizations

**Model deployment**
Will put the pre-trained models online and rewrite them for colab

**Keep up model development**
Research Reinforcement Learning for NLP

hai.bui
@me.com