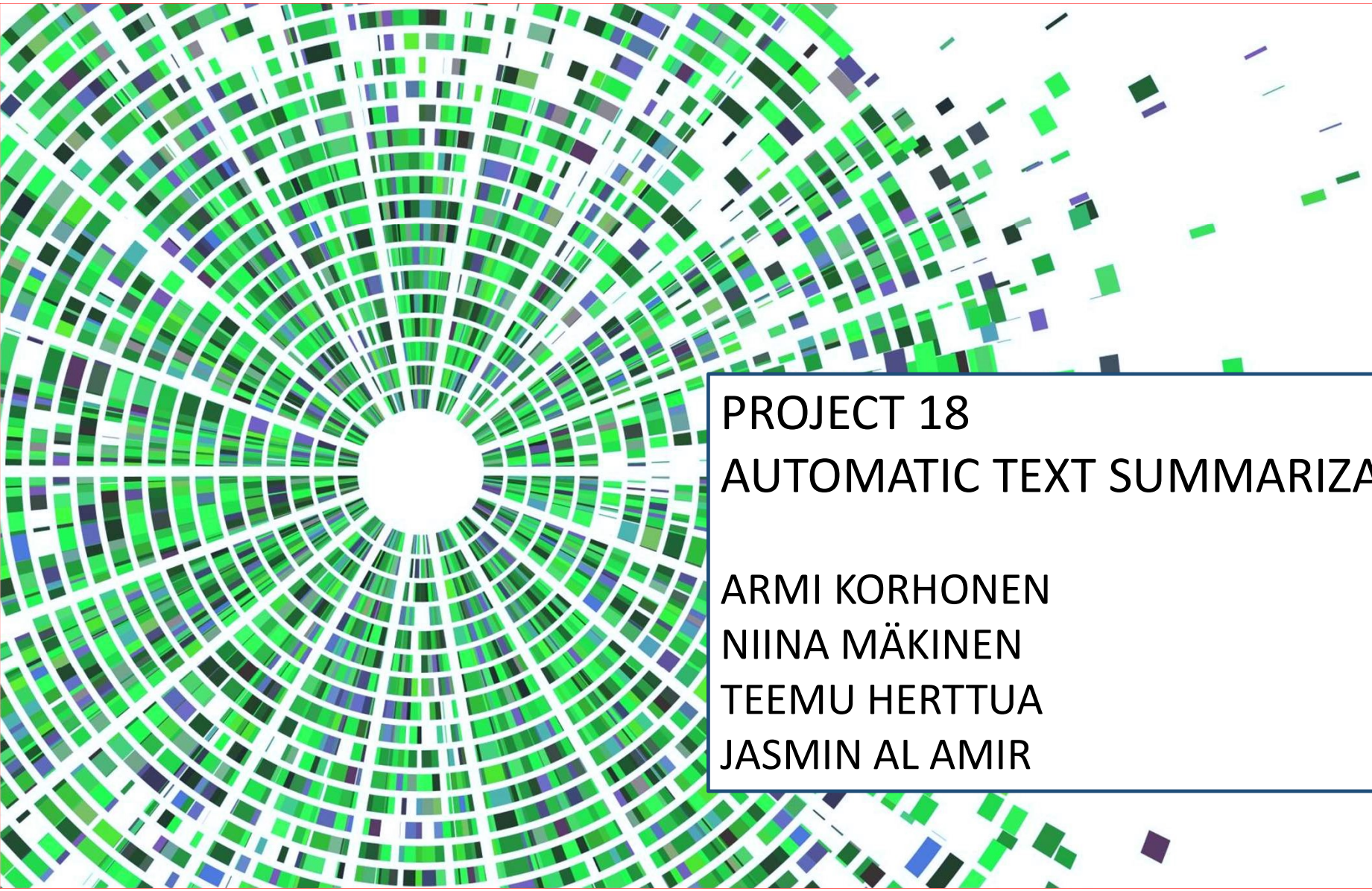


PROJECT 18
AUTOMATIC TEXT SUMMARIZATION

ARMI KORHONEN
NIINA MÄKINEN
TEEMU HERTTUA
JASMIN AL AMIR

ABSTRACT

This project aims to implement new approaches for automatic text summarization and evaluate their performances on small sample dataset.

As a benchmark, four different summarization algorithms are implemented, tested for performance and evaluated using the standard ROUGE-2 and ROUGE-3 evaluations.

Algorithms are further developed with the help of Spacy named entity recognition. The developed new algorithm is then implemented, tested for performance and evaluated against the existing algorithms.

TOPICS INVESTIGATED

- Summarization algorithms
- ROUGE-evaluation algorithm
- HTML news article parsing techniques
- Python GUI development
- Named entity recognition as basis for summarization algorithms

GROUP RESPONSIBILITIES

Niina Mäkinen – Documentation

Jasmin Al Amir – Planning, Documentation

Teemu Herttua – GUI, HTML parsing

Armi Korhonen – ROUGE-evaluation

In addition, everyone will have plenty more to do!

DATA SOURCES

- During development and testing: CNN/DailyMail dataset
- Currently: DailyMail.co.uk as a source for articles, English language only
- In future: Other data sources? RSS feed with news article URLs and summaries? More languages?

TECHNOLOGIES AND TOOLS USED

- Python 3.8
- PyTLDR fork that works with Python 3.8
  - Latent Semantic Analysis by Ozsoy
  - Latent Semantic Analysis by Steinberger
  - Relevance
  - TextRank
- Tkinter, Beautifulsoup, Newspaper and Spacy (with pretrained model en_core_web_sm), amongst others

IMPLEMENTATION - EXAMPLE

- URL: http://www.dailymail.co.uk/news/article-8865747/Rishi-Sunaks-new-lockdown-millions-Chancellor-prepares-rescue-deal-companies.html
- 4 different summaries with 5 sentences created using PyTLDR
  - Picked for example: Latent Semantic Analysis by Ozsoy

IMPLEMENTATION – AUTOMATIC TEXT SUMMARY

The Chancellor will update MPs this morning on a package set to cost hundreds of millions of pounds to help firms in sectors such as hospitality, which have been badly hit by new restrictions this month.

Office for National Statistics data revealed ministers have borrowed £208billion over the six months since April, working out at around £1.14billion daily.

The Chancellor (pictured in Downing Street yesterday) will update MPs this morning on a package of support set to cost hundreds of millions of pounds to help firms in sectors such as hospitality, which have been badly hit by new restrictions this month

Another £36.1billion was borrowed in September - the third-highest month on record and compared to just £7billion a year ago - as tax revenues slumped and the Treasury poured out bailout money

Government sources played down reports that chief medical officer Professor Chris Whitty is drawing up plans for a series of regional 'circuit breakers' next month, with even tighter restrictions.

IMPLEMENTATION – REFERENCE TEXT

Job rescue scheme 'to be extended to London and other Tier 2 lockdown areas' TODAY: Rishi Sunak will unveil new subsidies as government borrowing balloons to £1bn a DAY

- **Chancellor will update MPs this morning on a package of support set to cost hundreds of millions of pounds**
- **He prepared to reach deeper into Treasury's coffers as figures showed Government's daily £1bn borrowing**
- **Efforts to bring Government spending under control have been derailed by a second surge of coronavirus**
- **Ministers announced yesterday that South Yorkshire would be moved into the top Tier Three restrictions**

By JASON GROVES POLITICAL EDITOR FOR THE DAILY MAIL and JACK ELSOM FOR MAILONLINE
PUBLISHED: 22:47 BST, 21 October 2020 | **UPDATED:** 03:45 BST, 22 October 2020

Reference text for ROUGE evaluation

IMPLEMENTATION – ROUGE SCORES

- Using the reference text from the human written summary and the LSA Ozsay algorithmic summary for ROUGE scores
  - ROUGE-2 recall: 0.3684210526315789 -> **36.8%**
  - ROUGE-2 precision: 0.2160493827160493 7 -> **21.6%**
  - ROUGE-3 recall: 0.3085106382978723 4 -> **30.9%**
  - ROUGE-3 precision: 0.1790123456790123 4 -> **17.9%**

- Results could be better
- Further data and analysis needed

IMPLEMENTATION – NAMED ENTITIES

this morning TIME
hundreds of millions of pounds MONEY
this month DATE
National Statistics ORG
the six months since April DATE
around £1.14billion MONEY
daily DATE
Downing Street FAC
yesterday DATE

this morning TIME
hundreds of millions of pounds MONEY
this month DATE
36.1billion MONEY
September - the third-highest month DATE
just £7billion MONEY
a year ago DATE
Treasury ORG
Chris Whitty PERSON
next month DATE

- Name of Rishi Sunak is missing from the summary, even though he's in the headline (referred to as "The Chancellor" in the summary)
- Next step: improve the algorithm with named entities

IMPLEMENTATION TIMELINE

4 people working simultaneously

Documentation

Iterative Script & Algorithm Development

Research

PROJECT TASKS

- Documentation (ongoing)
- Research and analysis (ongoing)
- PyTLDR summarization (done)
- ROUGE evaluation (done)
- HTML parsing (done)
- Spacy named-entities (done)
- Seminar presentation (done)
- GUI development (mostly done)
- Summarization algorithm using NER (TODO)
- Performance metrics (TODO)
- Putting it all together (TODO)
- **We expect to finish the project on time!**

FURTHER DEVELOPMENT

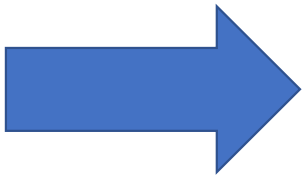
- Other data sources?
- Other languages?
- Hosted summarization application with a web interface?

FURTHER DEVELOPMENT – ALERT SUMMARIES?

- Utilizing and customizing Spacy named entities in summaries
- Connecting to Google Alerts

Super cells of rain south west of Alice
Alice Springs News Online
"Last night, however, isolated severe thunderstorm **supercells** caused heavy rain and possibly hail in the Lasseter district." They are shown as rainbow ...

Mobile Games Market 2020 New Trends, Opportunities After COVID-19 & Major Companies ...
PRnews Leader
Mobile Games Market 2020 New Trends, Opportunities After COVID-19 & Major Companies – **Supercell** Oy, Gameloft, Glu Mobile Inc., Zynga Inc., The ...



Document text summary with named entity of interest (e.g. ”Supercell”, ”ORG”)

Thank you!
Questions?

arkorhon20@student.oulu.fi