S7 Group Project - Newsjam - First Report

Shane Kaszefski-Yaschuk Joseph Keenan Adelia Khasanova
Maxime Méloux Mahamadi Nikiema

NLP Master – Université de Lorraine

October 1, 2021

1 Motivation

The ongoing COVID-19 pandemic has increased the average daily time spent on social media by 65% for U.S. citizens[1]. In particular, news website and social media are the main sources of updates related to the pandemic[2]. However, 72% of U.S. adults report experiencing news fatigue, with young adults being among the most affected[3]. We believe that there is a need for a centralized, factual and concise news source that could enable its users to stay informed about the current situation while requiring little time or emotional involvement.

2 Proposal

Our goal is to create a system (*Newsjam*) that is able to scrape and summarize online French news articles concerning the COVID-19 pandemic in France. We intend to create a bot that will post the summarized articles on Twitter. Our vision of the final product encompasses the following elements:

- Corpus building: Due to the specific topic, language and geographical area we are focusing on, we expect that we'll need to create our own corpus. This corpus will be created by scraping online French news articles about the pandemic in France and possibly annotating them with a reference summary (depending on the chosen summarization evaluation method).

- Scraping: When provided with one or several links to appropriate news articles, the system will be able to automatically scrape.

- Summarization: The scraped text will be tokenized and fed to a model that generates a short summary of the article. The output summary should fit within one tweet.

- Posting: The summarized text will be posted on Twitter automatically.

- Regular updates: The system will regularly post a condensed breakdown of the current COVID-19 situation in France on Twitter. This breakdown will contain statistics such as the number of new cases detected within the last day or week.

In addition, we propose a number of future potential extensions to increase the scope of the project:

- Article selection: The system should be able to automatically find relevant articles online and to rank them by order of importance, so that the summarized news do not overlap and only include important information.

- Topic generalization: The system can be expanded to work for other short to medium-term events or topics such as a particular sport competition or elections in a given country.

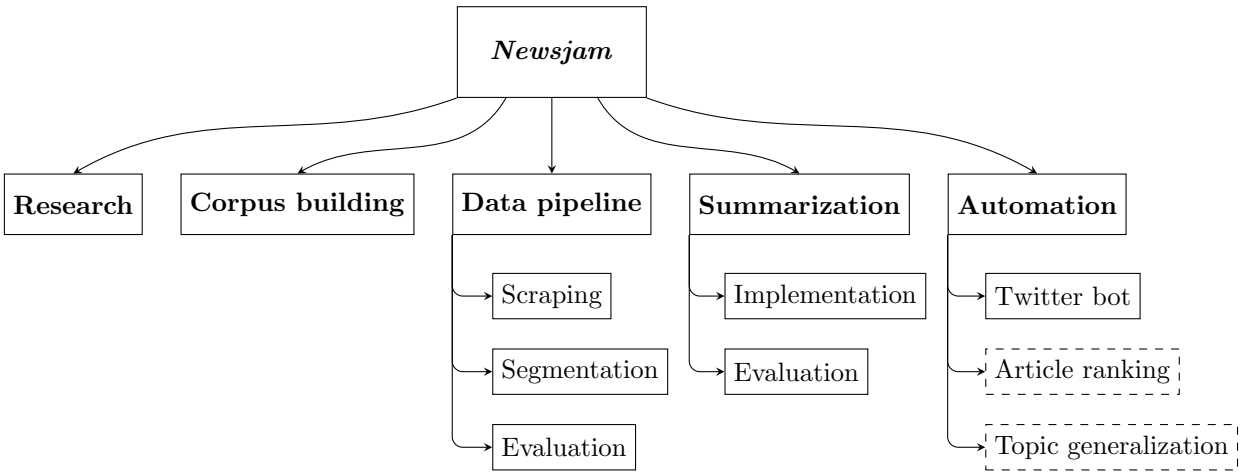
[1] Statista Research Department (2021), Additional daily time spent on social media platforms by users in the United States due to coronavirus pandemic as of March 2020, Statista

[2] Nielsen et al. (2021), An Ongoing Infodemic: How People in Eight Countries Access and Rate News and Information about Coronavirus a Year into the Pandemic, Reuters Institute

[3] Watson, A. (2020), Share of adults who feel overwhelmed by the amount of information in coronavirus news coverage in the United States as of April 2020, by age group, Statista

3 Project overview

The following overview gives a high-level overview of *Newsjam* from a project management point of view. The project is divided into 5 Work Packages (WP):



4 Evaluation

Throughout our project we expect to perform two main evaluation tasks:

• Scraping: Since article scraping and text retrieval from a news source (URL) is to be automated, we will need to evaluate the quality of the retrieved text compared to a gold standard. The scraper can then be incrementally improved by removing advertisements, navigation elements and other irrelevant data from the output. We will need to come up with an appropriate metric for that purpose.

• Summarization: To evaluate the quality of the summarization component, we are currently considering two different approaches. The first one involves annotating articles in our corpus with human-generated reference summaries and using existing metrics such as BLEU or ROUGE to measure the performance of our model. This has the advantage of being easier to implement, but is inherently flawed due to the difficulty of writing a gold standard summary for a given article. The second approach involves comparing the full-length article text with the summary directly. This can be done using processes such as *latent semantic analysis*. This approach is harder to implement but leads to a more accurate evaluation.

5 Organization

Our team is made up of various members with diverse backgrounds:

- Shane Kaszefski-Yaschuk has experience in both linguistics and computer science.
- Joseph Keenan has experience in linguistics as well as a base knowledge of computer science.
- Adelia Khasanova has a linguistic background and a good grasp of data analysis with Python.
- Maxime Méloux has experience in applied mathematics and computer science for NLP.
- Mahamadi Nikiema has experience in applied mathematics and statistics.

We will be using the Scrum/Agile framework to ensure proper project management.