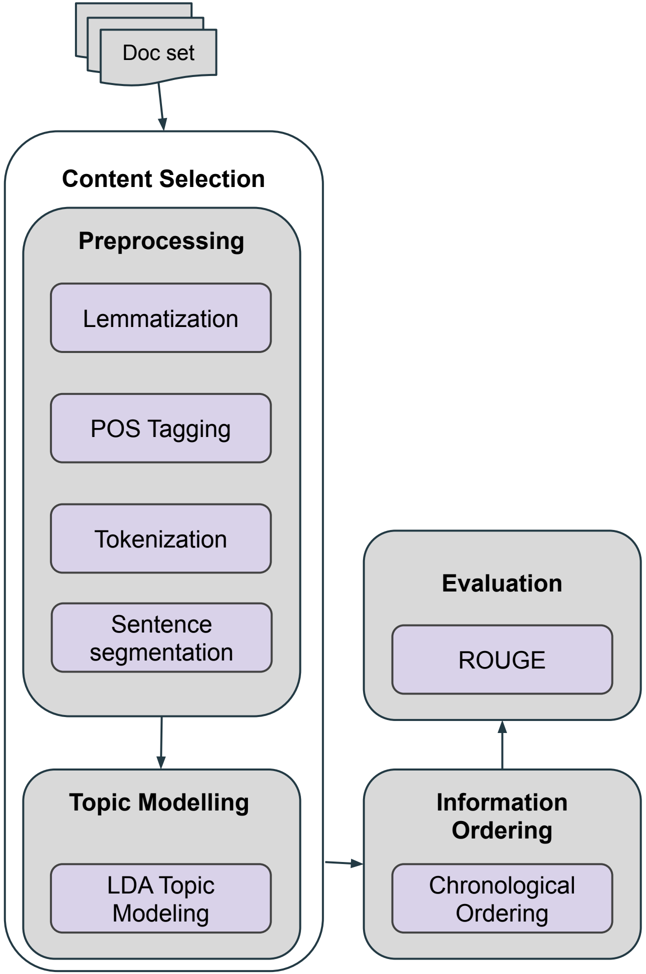
Deliverable #2

Multi-document Summarization System
LING 573

Group 1 - Elena Khasanova, Erica Gardner, Saumya Shah, Sophia Chan and Vikash Kumar

System Overview

1. **Data Extraction** - Arrange documents by topic id and extract documents in XML format
2. **Content Selection**
   a. Preprocessing - Stop word removal, lemmatization, sentence segmentation and tokenization
   b. LDA - Produce a set of topics over document and a vocabulary over the topics
3. **Information Ordering -** Pick the top sentences from the LDA module for the summary
4. **Content Realization -** Removing parentheses, removing adverbs, eliminating sentences shorter than 8 words



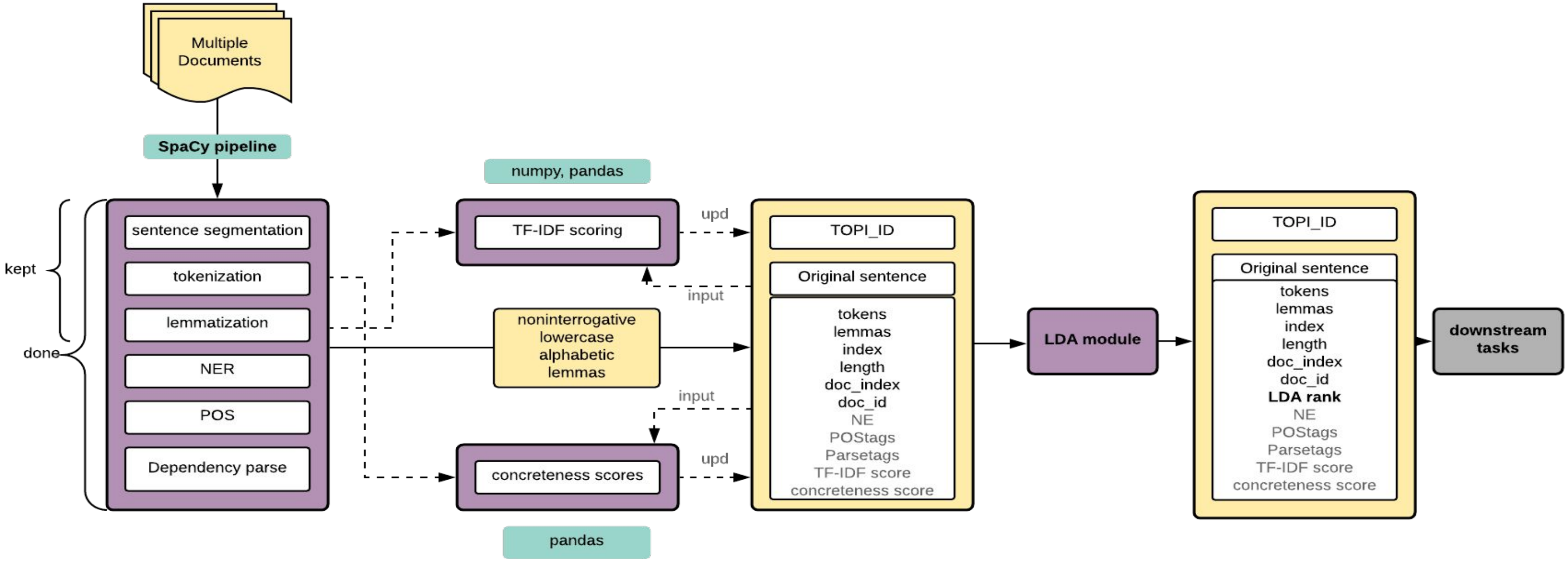
Literature Review

The methods we came across that perform extractive summarization are given below:

1. Using **word-frequency algorithms** to choose the sentences to be included in the summary
2. **TextRank matrix** that shows the **cosine similarity** between sentences and use of universal sentence embeddings to generate the top N sentences that will be used for the summary
3. The use of **skip-thought vectors** for sentence similarity and clustering to find topic clusters within a sentence

For all these methods, we found that a ranking component is essential. And since this method is predominantly unsupervised, an unsupervised technique to find different topics/clusters within a document is useful to identify the important aspects of the text.

Preprocessing



Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014).Concreteness ratings for 40 thousand generally known English word lemmas.Behavior Research Methods, 46, 904-911.

Future steps in preprocessing

- **TF-IDF scoring** :
  - sent.TF-IDF = sum (lemma.TF-IDF)
- **Concreteness scoring:**
  - sent.C = sum (lemma.C) if lemma in *concreteness_df*
- **Overall sentence score:**
  - a * sent.LDAscore + b * sent.TF-IDF + c * sent.C; a, b, c - estimated coeff
  - sent.LDAscore*sent.TF-IDF*sent.C, rescaled from 0 to 1
- **Input delexicalization** :
  - replace **x** with **y** if *sim(x, y)* > *thres;* pick *thres*
- **Skip-thought vectors**

Content Selection - Topic Modelling

- **Latent Dirichlet Allocation(LDA)** is a popular algorithm to estimate latent/hidden topics in any given set of document.
- It treats the corpus as **bag of words** and outputs two probability distributions :
  - Distribution of topics over documents
  - Distribution of words over topics
- We used **Genism’s** multi core implementation

LDA Parameters

- **Number of topics:**
    - Hyperparameter - Limitation of the model
    - Difficult to evaluate intrinsically
    - Topic coherence can be used to prune out irrelevant topics
    - 3 topics for each document set for D2
- **Alpha**
    - This prior intuitively controls the **distribution of topics over the documents**. A higher alpha value means that each document is likely to contain mixture of all the topics, a low alpha value gives preference to a distribution where documents belong to a single topic
    - Set to **1/{num of topics} = 0.33**
- **Beta**
    - This prior controls for the **distribution of words among topics**. A higher beta-value yields overlapping topics while a low beta value one word to strongly belong to multiple topics.

Sample topic

```
[(0, '0.028*"nepal" + 0.021*"say" + 0.019*"king" + 0.018*"gyanendra" +
0.014*"government" + 0.012*"democracy" + 0.010*"minister" + 0.008*"prime" +
0.008*"china" + 0.008*"new"'),

(1, '0.007*"report" + 0.006*"peace" + 0.005*"news" + 0.005*"street" +
0.005*"katmandu" + 0.004*"deuba" + 0.004*"calm" + 0.004*"china" + 0.004*"unrest" +
0.004*"colleague"'),

(2, '0.005*"non" + 0.004*"state" + 0.003*"nepal" + 0.003*"principle" + 0.003*"police" +
0.003*"violence" + 0.003*"end" + 0.003*"patrol" + 0.003*"launch" + 0.003*"estimate"')]
```

Information Ordering

- Determined sentence order based on LDA score
  - LDA score output had the **top 5 sentences**
  - Added these to summary in order of rank

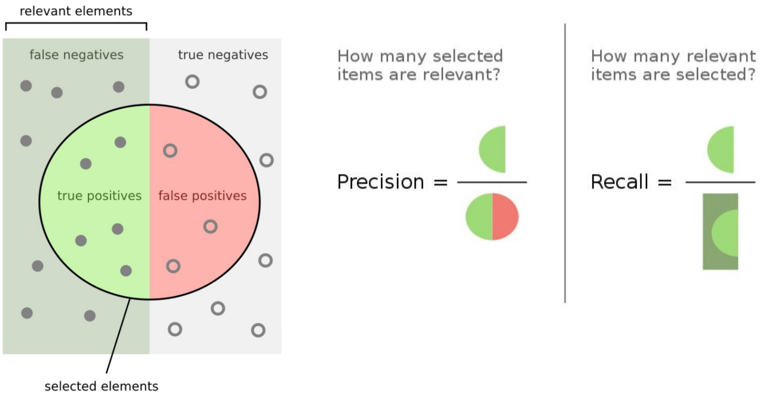
- To do for next deliverable:
  - Adjust approach to place more emphasis on **cohesion, coherence**
  - Incorporate chronology by considering **document time stamp**
  - Incorporate topicality by using **cosine similarity** to add sentences most similar to those already in summary

Content Realization

- Incorporated several basic heuristics to cut down on irrelevant information:
  a. Did not include sentences with **< 8 words**
  b. Removed some parenthetical expressions, those with **(), []**
  c. Removed adverbs
  d. If output was less than 100 words, **added a sentence that would bring it as close to 100 words** as possible, even if it wasn’t the next highly ranked sentence
- For next time:
  a. Incorporate additional heuristics
  b. Exploring **learning or phrase structure-based** approaches

Results: ROUGE-1 and ROUGE-2

- **ROUGE-1**
    - Average Recall: **0.18334**
    - Average Precision: **0.25957**
    - Average F-1: **0.21316**
- **ROUGE-2**
    - Average Recall: **0.04340**
    - Average Precision: **0.06116**
    - Average F-1: **0.05031**



$$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} \right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Results: Comparison to baselines

	ROUGE-2 (Average F)
D2 System	0.05031 (95% CI [0.04008, 0.06069])
LEAD Baseline	0.05376
MEAD Baseline	0.05927

Error Analysis

D1002

Some of the police officers who shot and killed Amadou Diallo have told associates that their attention was drawn to Diallo when they saw him standing on the stoop of a Bronx apartment building and thought they saw him peering into the window of a first-floor apartment~~, according to people with knowledge of the case.~~

D1014

Many spoiled teenagers from one-child families eat a particular kind of food, such as meat or sweets, which might lead them to develop **obesity**, diabetes or heart disease.

The study shows that the **12 percent obesity** is to the standard in developed countries.

The research showed that the ratio of adults suffering from **obesity** has reached **12.6 percent**, and that of youngsters is 11.35 percent.

Error Analysis

- Why did it happen?
    - **Attribution**: no filtering mechanism
    - **Redundancy**
        - No metric to measure similarity between sentences
        - Lost mapping to LDA topics and documents during content selection
    - Pipelines are difficult
        - Errors can propagate from one layer to the next

Potential Solutions

- What can be changed?
  - Preserve information about LDA topics and which document the sentence came from
    - Use **date of publication**
    - Distribute sentences across subtopics
  - Prevent against summary with two sentences with **high cosine similarity**
  - Use a heuristic to remove attribution
    - e.g. Remove portion that contains **attributive word** + occurs after comma and before period

Results: Attribution

D1002

Some of the police officers who shot and killed Amadou Diallo have told associates that their attention was drawn to Diallo when they saw him standing on the stoop of a Bronx apartment building and thought they saw him peering into the window of a first-floor apartment~~, according to people with knowledge of the case~~.

Results: Redundancy

D1014

Many spoiled teenagers from one-child families eat a particular kind of food, such as meat or sweets, which might lead them to develop**obesity**, diabetes or heart disease.

The study shows that the **12 percent** **obesity** is to the standard in developed countries.

The research showed that the ratio of adults suffering from**obesity** has reached **12.6 percent**, and that of youngsters is 11.35 percent.

For the next deliverable

- **Content Selection**
  - We are aiming to experiment with **hyper parameters** in LDA and also augment LDA topic modeling with **TF-IDF** and **concreteness** scoring
  - Delexicalization of the input based on word vector similarity.
  - Explore **skip-thought vectors** and clustering methods
- **Information ordering**
  - We plan to use the document **meta-information**
  - Experiment with **chronological ordering**
  - Promote **coherence** using cosine similarity.
- **Content Realization**
  - Use phrase structure to ensure **cohesion**
  - Experiment with **learning methods**

Conclusion

- We largely focused working out the system architecture and ensuring that the separate components run smoothly following one another. We also experimented with several content selection methods and built the scaffolding for a more refined versions of this module.

- The end-to-end system achieved the ROUGE-1 and ROUGE-2 scores comparable with the baselines and outperformed the baseline in one out of three cases. The scores show that our approach has the potential to yield good results.

Thank you!