CS 267 Homework1 Part 2 Report

I Human Experiment

For the three webpages, yahoo, IMDB and Wikipedia Canada page, we summary what information we expect to be able to find in descending order of importance and the query we may input into a search engine for each of these information needs in the following table:

Web Page	Information Need	Query
Yahoo Landing Page	Top news on a specific topic	1 “Trump news Yahoo”
	Top news on a specific category	2 “Sports news Yahoo”
	Top news occurred or associated with a specific place	3 “San Jose news Yahoo”
nternet Movie Database Landing Page	Name of new movie releases	4 “new movies”
	Review of a specific movie	5 “The Lion King 2019 review”
	Trailer of a specific movie	6 “The Lion King 2019 Trailer”
Wikipedia Canada page	Total area of Canada	7 “Canada total area”
	Population of Canada	8 “Canada population”
	Language spoken in Canada	9 “Canada language”

Next , we make a summary of each of the 3 pages:

- Yahoo Landing Page:
  yahoo home page, yahoo search, yahoo mail, yahoo messenger, yahoo games, news, sports: NBC Sports BayArea, Yahoo Sports, US: people, yahoo style UK. celebrity, lifestyle, style, world, entertainment, science. News, email and search are just the beginning. Discover more every day. Find your yodel.
  In the summary there appeared: 2 words relevant to query 1, all words relevant to query 2, and 2 words relevant to query 3. In total the summary contained all query words once.
- IMDB Landing Page:
  IMDb is the world's most popular and authoritative source for movie, TV and celebrity content. Find ratings and reviews for the newest movie and TV shows. Ratings and Reviews for New Movies and TV Shows, movies, films, movie database, actors, actresses, directors, hollywood, stars, quotes
  In the summary there appeared: all words relevant to query 4, 1 word relevant to query 5, and 0 word relevant to query 6. In total the summary contained all query words once
- Wikipedia Canada Page:
  Canada, country in the northern part of north America extending from the Atlantic to the Pacific and northward into the Arctic Ocean, covering 9.98 million square kilometres (3.85 million square miles), making it the world's second-largest country by total area. Canada’s population reached 37,602,103 in 2019. Canada's capital is Ottawa, and its three largest metropolitan areas are Toronto, Montreal, and Vancouver. Etymology, History, Indigenous peoples, European colonization, Confederation and expansion, Early 20th century, Contemporary era, Geography and climate, Government and politics, Law, Foreign relations and military, Provinces and territories, Economy,

Science and technology, Demographics, Health, Education, Ethnicity, Religion, Languages, Culture, Symbols, Literature, Visual arts, Music, Sports

In the summary there appeared: all words relevant to query 7, all word relevant to query 8, and all word relevant to query 9. In total the summary did contain all query words three times.

II Yioop Experiment

In each experiment the Max Page Summary Length was set to only 2000. By entering the webpage url after selecting test page by Uri under the 'Test Options', we can get certain test results including [Description] and [Description_Scores] useful for obtaining summary of the page, which looks like figure 1.

i) Yahoo Landing Page:

Byte range to download was kept as 100,000 under basic summarizer by default, and set to 500,000 under Centroid Weighted summarizer. First 4 sentences of summary returned by both summarizers are identical, but the Centroid Weighted summarizer gives more detailed 5th sentence, so we adopt the top 5 sentences from the Centroid Weighted summarizer:

```
News, email and search are just the beginning. Discover more every
day. Find your yodel. Yahoo.  Sign in. Mail Sign in to view your
mail.
```

In the summary there appeared: 2 words relevant to query 1, 2 words relevant to query 2, and 2 words relevant to query 3. In total the summary never contained all query words.

```
    [DESCRIPTION] => .. News, email and search are just the beginning.
Discover more every day. Find your yodel.  Yahoo .  Mail Sign in to view
your mail.  News Finance Sports Politics Entertainment Lifestyle More.
    [WORD_CLOUD] => Array
        (
            [0] => yahoo
            [1] => news
            [2] => search
            [3] => email
            [4] => sign
        )

    [DESCRIPTION_SCORES] => Array
        (
            [1] => 0.7247346681798889
            [9] => 0.347728218135868
            [16] => 0.4747366782667451
            [17] => 0.3058614591867233
            [24] => 0.1923680307423487
            [31] => 0
        )
```

Figure 1 Screen Shot of description and description_score portion of the test result on Yahoo landing page (under Centroid Weighted summarizer)

ii) IMDB landing page

Byte range to download was set as 50,000 under both summarizers. Figure 2 below shows the description scores returned by using basic and Centroid Weighted summarizer. In comparison, the Centroid Weighted summarizer assigned the first few arrays, mostly below 67, a higher description score.

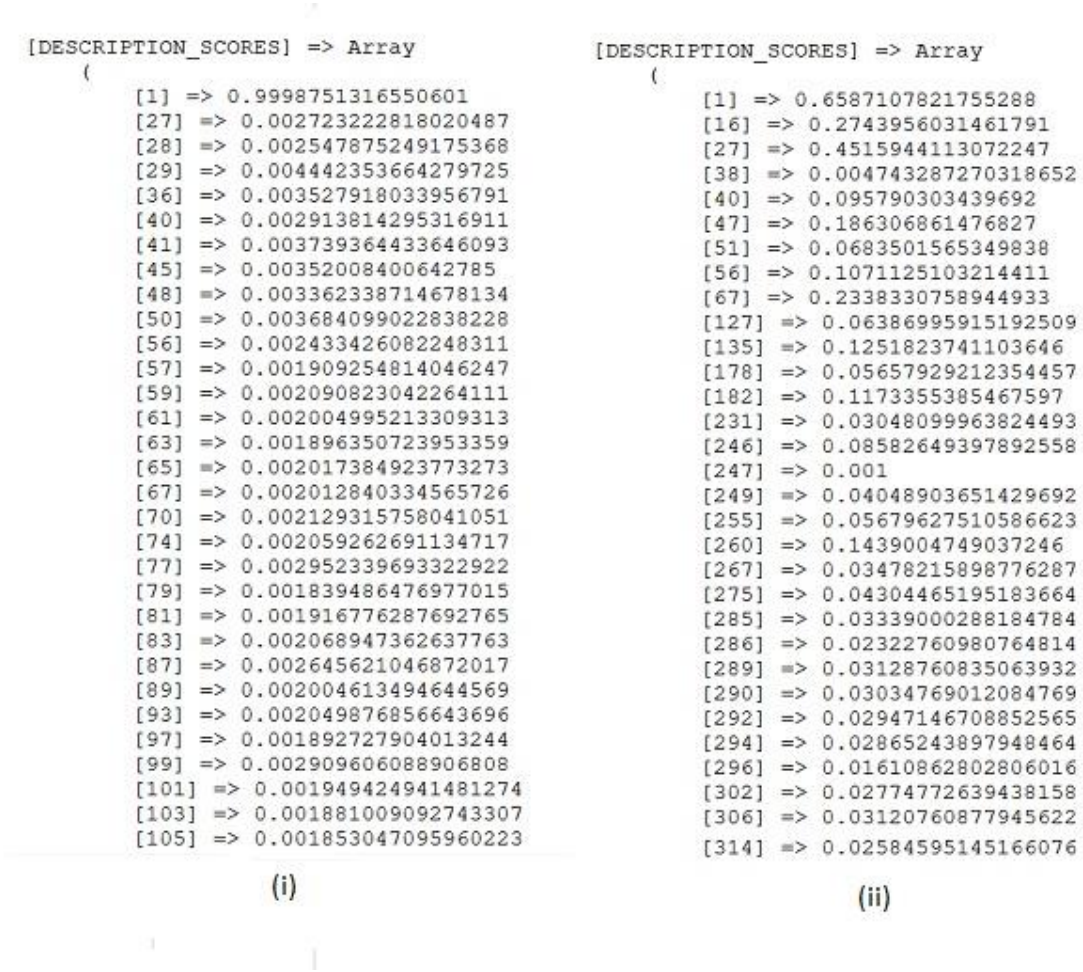

Figure 2 Screen Shot of description_score returned for IMDB landing page by (i) BASIC summarizer and (ii) Centroid Weighted summarizer

The top 5 sentences produced by Centroid Weighted summarizer:

IMDb is the world's most popular and authoritative source for movie, TV and celebrity content. Find ratings and reviews for the newest movie and TV shows. Ratings and Reviews for New Movies and TV Shows - IMDb. All Titles TV Episodes Names Companies Keywords. Watch Now For Free.

In the summary there appeared: all words relevant to query 4, 1 word relevant to query 5, and 0 word relevant to query 6. In total the summary contained all query words once out of 3 times.

iii) Wikipedia Canada Page

Again, Byte range to download was set as 50,000 under both the basic and Centroid Weighted summarizer.
The top five sentences produced by Centroid Weighted summarizer:

Country in North America. Coordinates: 60°N 95°W / 60°N 95°W / 60; -95. Canada is a country in the northern part of North America. Its ten provinces and three territories extend from the Atlantic to the Pacific and northward into the Arctic Ocean, covering 9.98 million square kilometres (3.85 million square miles), making it the world's second-largest country by total area. Its southern border with the United States, stretching some 8,891 kilometres (5,525 mi), is the world's longest bi-national land border. Canada's capital is Ottawa, and its three largest metropolitan areas are Toronto, Montreal, and Vancouver.

> In the summary there appeared: all words relevant to query 7, 1 word relevant to query 8, and 1 word relevant to query 9. In total the summary contained all query words once.

In comparison, both human-generated and Yioop generated summaries contain exactly the same sentences as wrapped in a tag <meta name="description" content= “”> that can be located in the webpage’s source html file. This indicates that the description meta data of the webpage can serve as an excellent reference for the summary of the page. However, the human generated summaries can contain lots of nitty gritty details regarding the webpage, many of which can easily match the potential query, while Yioop generated top-5 sentence summary is based on a description score measuring relevance of terms extracted from the webpage with the overall content of the webpage. Hence the chance maybe lower that the entire query term appears in the Yioop generated summary.