Search Engine Technology, Software Implementation Project Report
Name: Rupayan Basu UNI: rb3034

GISTIFY - A Software For Automatic Text Summarization Using the Vertex Cover Algorithm

1. Introduction

Automatic text summarization is a key area of research in NLP and IR that seeks to help us deal with vast quantities of information that are increasingly available by providing minimal yet meaningful summaries. An example is in the news domain where summaries of various news articles could help time-pressed readers get a gist of the main events, or lead them to articles that are of particular interest to them.

There are many approaches to summarization - *extractive* summarization picks key sentences from an article to produce a summary, while *abstractive* summaries construct new sentences to capture the main ideas in an article. *Single document summarization* and *multi document summarization* are examples of other approaches. Gistify uses the extractive, single document approach.

2. Method and Justification

In gistify's approach to summarization, the key question that is answered by the summary is - "What is the article about?" In order to answer this question, gistify needs to select the sentences in the article that are most representative of the overall content of the article. Hence, extraneous details are discarded (even if they might be important on their own), while the sentences that tell us most about the overall meaning of the article are retained.

The method followed to achieve this involves representing the entire text article as an undirected graph as follows: sentences in an article are represented as nodes and edges between nodes indicate that the two nodes are 'similar', that is, they have a certain number (above a fixed threshold) of words in common. A minimum vertex cover of this graph produces a set of nodes (sentences) such that each edge in the graph is incident to at least one node in the set. This set of nodes then gives us a summary of the article.

As a justification of this method, let us first consider 'similairy' between sentences. Here, we are *not* considering semantic similarity between sentences, but are using a method akin to the 'bag of words' approach, where if two sentences have a certain number of words in common, then they are deemed similar. For example: 'John runs faster than Mary' is not semantically similar to 'Mary runs faster than John'. However, since both sentences have the same words, we can say that the two sentences are about the same thing, that is, the relative running speeds of John and Mary. Hence, a summary should ideally contain either of the two, in order to give the reader an idea of what the original article is about.

Finally, the 2-approximation vertex cover algorithm is used to select a subset of vertices since the minimum vertex cover problem is NP-complete. A vertex cover is a subset of vertices such that every edge is incident on at least one vertex in the subset. The intuition behind using the algorithm is as follows: since edges represent 'content similarity' between sentences, a vertex cover would then have 'covered' all of the content in an article. Vertices that are not connected to the graph are discarded since they represent isolated details that, although potentially informative, are not representative of the overall meaning of the article.

3. Running Gistify

To produce the summary of a text article in sample.txt, run the command below. Gistify produces the summary of the article as well as the counts:

$./gistify -i input_files/sample2.txt

Emperor of Japan Akihito accompanied by Empress Michiko arrived in New Delhi on Saturday for a six-day visit described by the Government as “one of the biggest moments in India’s diplomatic engagement this year’’.
It is the first time that the Emperor and Empress of Japan are coming to India and it is also a first that India has hosted the same two dignitaries on a state visit after a lapse of 50 years.

Number of sentences in article: 5
Number of sentences in summary: 2

To see the original text article with the sentences selected for the summary highlighted in blue, run the following:

$./gistify -i input_files/sample2.txt -p

Emperor of Japan Akihito accompanied by Empress Michiko arrived in New Delhi on Saturday for a six-day visit described by the Government as “one of the biggest moments in India’s diplomatic engagement this year’’.
This was reflected in Prime Minister Manmohan Singh and his wife receiving the Japanese Royal couple at the airport.
This was a rare gesture Dr. Singh had reserved for US President Barack Obama in 2010 and his predecessor George Bush in 2006.
It is the first time that the Emperor and Empress of Japan are coming to India and it is also a first that India has hosted the same two dignitaries on a state visit after a lapse of 50 years.
“It has never happened before in the history of independent India,’’ pointed out Ministry of External Affairs spokesperson Syed Akbaruddin.
Number of sentences in article: 5
Number of sentences in summary: 2

Gistify uses a default threshold value of 3 to determine similarity between two sentences. This threshold can be changed with the -s option as follows:

$./gistify -i input_files/sample4.txt -s 4

Early on Sunday the spacecraft fired its main engine for more than 20 minutes, giving it the correct velocity to leave Earth's orbit.
The Mars Orbiter Mission (MOM), also known as Mangalyaan, is designed to demonstrate the technological capability to reach Mars orbit.
MOM tweeted: "Earth orbiting phase of the #Mangalyaan ended and now is on a course to encounter Mars after a journey of about 10 months around the Sun".
So engineers opted for a method of travel called a Hohmann Transfer Orbit to propel the spacecraft from Earth to Mars with the least amount of fuel possible.

Number of sentences in article: 17
Number of sentences in summary: 4

To compare gistify's summary with a gold standard summary, use the -t <testfile> option as follows:
$./gistify -i input_files/sample5.txt -t test_files/test5.txt
Google celebrates the 180th birth anniversary of Cuban physician and scientist Carlos Juan Finlay through a doodle on Tuesday.
Google pays honour to the man who propounded the path-breaking theory that yellow fever was spread by mosquitoes.
It has Carlos Juan Finlay's face amidst stagnant water, leaves and mosquitoes breeding on them.
This doodle is a tribute to Carlos Juan Finlay's theory and study which continues to save many lives today.
In 1879, he was appointed by the Cuban government to work with a North American commission studying the causes of yellow fever.
After two years, Carlos Juan Finlay was sent as the Cuban delegate to the fifth International Sanitary Conference in Washington DC.
At the conference, he urged those present to study yellow fever vectors.
Finlay later theorised that the carrier of yellow fever was the mosquito Culex fasciatus,now known as Aedes aegypti.

Number of sentences in article: 15
Number of sentences in summary: 8

Precision: 0.50
Recall: 1.00

4. Implementation details

Language used: Python 2.7.3
Modules/packages: nltk, re, sys, getopt, colorama and the nltk corpus for stopwords
Source file: gistify.py

Preprocessing:
Preprocessing is a crucial first step, since it allows precise word based comparison of sentences when testing for sentence similarity. The different steps are:
-To prepare the text for tokenization, regular expressions were used to remove special characters, convert subtitles to sentences, and placing quotes before periods rather than after, that is, converting ." to ". so that all sentences ended with the period.
-The punkt tokenizer was then used to separate the text into sentences. A tokenizer was used because simply splitting the text using '.' as a separator would cause problems with words like U.S.A and Mr. etc.
-stop words such as 'a', 'of', 'with' and others were removed to retain the most significant words.
-Then to help with word-word comparisons, the sentences were lower cased.
-Lemmatization was used to allow words such as 'good' and 'better' to be identified as similar words.
-Finally, stemming was done to reduce words to their roots to allow different inflections of words to be identified as similar words. Example, walk, walks, walked and walking are all similar words.

API details:
3.1 get_content(infile) - returns entire body of text from input file.
3.2 get_sentences(content) - takes a bulk of text and returns list of (id,sentence) pairs called sentuples. Performs minor preprocessing before using the punkt tokenizer.
3.3 preprocess_sentences(sentuples) - takes list of (id, sentence) and returns list of (id, preprocessed_sentence). Removes all special characters other than letters, digits and whitespaces. Removes stop words, and performs lower casing, stemming and lemmatization.
3.4 find_edges(sentuples, threshold): - converts sentences to an undirected graph. Takes list of (id, preprocessed_sentence) and an integer threshold (default=3), returns list of (id1, id2) representing endges between nodes (sentences) id1 and id2.
3.5 get_vertex_cover(edges) - Implements a linear time approximation of the Minimum Vertex Cover algorithm Vertex cover has size <= 2 * minimum size (optimal solution). Returns sorted resultant list of edges. Sorting takes O(nlogn) time, so total running time is O(nlogn).
3.6 get_orig_sentences(vc_edges, sentuples) - takes result of vertex cover and original list of (id, sentences) Returns list of (unpreprocessed) sentences selected for the summary.
3.7 convert_to_string(sentence_list) - Converts list of sentences to string for printing. Takes list of sentences selected for the summary and returns a string.
3.8 pretty_print(sentence_list, sentuples) - prints all sentences, with those in the summary highlighted in blue.
3.9 test(tfile, sentence_list) - compares the summary with a gold standard and prints precision and recall.

5. Evaluation and Testing

As gold standards for summarization of articles are hard to come by, I have used the gold standard summaries used in some research papers and articles on summarization, for evaluating gistify. The tests have shown higher recall values than precision, indicating that gistify selects the relevant sentences, but it also selects a few extra sentences. This is expected since gistify uses a 2-approximation minimum vertex cover algorithm, and hence it selects at most twice as many vertices as in the optimal solution.
On comparing the summaries produced by gistify with the gold standards, gistify produced precision and recall scores for each test. Two sample tests are provided below:

Test 1 - (gold std selected from the paper ‘An Evaluation Road Map for Summarization Research’)
$ ./gistify -i input_files/sample6.txt -t test_files/test6.txt -s 2 -p
US president Bill Clinton has arrived in Moscow for his first meeting with Russia's new president Vladmir Putin.
The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.
High on the agenda will be the United State's plans to build a missile shield in Alaska.
Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 which bans any anti-missile devices.
Clinton -- in his last few months of office and keen to make his mark in American history - will be seeking to secure some sort of concession from Putin.
The Russian leader has said that he will suggest an alternative to the US system.
Kremlin officials said Putin would propose a system that would shoot down the missiles with interceptors shortly after they were fired rather than high in their trajectory.
"We'll talk about it in Russia," Clinton told reporters before leaving Berlin for Moscow.
"It won't be long now".
Accompanying the President is US Secretary of State Madeline Albright.
"What's new is that Putin is signalling that he is open to discuss it, that he is ready for talks," she said.

"We will discuss it".
Arms control will not be the only potentially troublesome issue.
US National Security Adviser Sandy Berger said last week Clinton would raise human rights and press freedom.
Number of sentences in article: 14
Number of sentences in summary: 6
Precision: 0.67
Recall: 0.80

Test 2 (taken from 'Extraction-Based Single-Document Summarization Using Random Indexing')
$ ./gistify -i input_files/sample7.txt -t test_files/test7.txt -s 4 -p
A solar eclipse occurs when the Moon passes between Earth and the Sun, thereby totally or partially obscuring Earth's view of the Sun.
This configuration can only occur during a new moon, when the Sun and Moon are in conjunction as seen from the Earth.
In ancient times, and in some cultures today, solar eclipses are attributed to mythical properties.
Total solar eclipses can be frightening events for people unaware of their astronomical nature, as the Sun suddenly disappears in the middle of the day and the sky darkens in a matter of minutes.
However, the spiritual attribution of solar eclipses is now largely disregarded.
Total solar eclipses are very rare events for any given place on Earth because totality is only seen where the Moon's umbra touches the Earth's surface.
A total solar eclipse is a spectacular natural phenomenon and many people consider travel to remote locations in order to observe one.
The 1999 total eclipse in Europe, said by some to be the most-watched eclipse in human history, helped to increase public awareness of the phenomenon.
This was illustrated by the number of people willing to make the trip to witness the 2005 annular eclipse and the 2006 total eclipse.
The next solar eclipse takes place on September 11, 2007, while the next total solar eclipse will occur on August 1, 2008.

Number of sentences in article: 10
Number of sentences in summary: 4
Precision: 0.50
Recall: 0.40

6. Conclusion

While the precision and recall scores of gistify might not be impressive, it should be noted that the P&R approach to testing is not considered to be the best for summarization (From 'An Evaluation Road Map for Summarization Research'). Other methods such as Relative Utility method and the Pyramid method might be better suited, but require manually created summaries, which are not easily available online.

An improvement to gistify would be to implement a more efficient approximation to the minimum vertex cover problem, so that it produces subsets closer to the optimal solution.

7. References

-An Evaluation Road Map for Summarization Research
(Breck Baldwin, Robert Donaway, Eduard Hovy, Elizabeth Liddy, Inderjeet Mani, Daniel Marcu, Kathleen McKeown, Vibhu Mittal, Marc Moens, Dragomir Radev, Karen Sparck Jones, Beth Sundheim, Simone Teufel, Ralph Weischedel, Michael White)

-SIGIR tutorial on summarization, 2004 (Dragomir Radev)

-A Survey on Automatic Text Summarization (Dipanjan Das, Andre F.T. Martins)

-Extraction-Based Single-Document Summarization Using Random Indexing
(Niladri Chatterjee, Shiwali Mohan)