Use TextRank to Extract Most Important Sentences in Article

And Visualize the Internal Graph

Ceshine Lee Follow
Dec 9, 2018 · 5 min read

Photo Credit

I'm trying to build a NLP system that can automatically highlight the important part of an article to help people to read long articles. The common practice is to start with a simple baseline model that is useful enough, and then incrementally improves the performance. The TextRank algorithm[1], which I also used as a baseline in a text summarization system, is a natural fit to this task.

Available Implementations

There are multiple open-sourced Python implementations of TextRank algorithm, including ceteri/pytextrank, davidadamojr/TextRank, and summanlp/textrank. They all come with different flavors of text pre-processings (for example, PyTextRank uses parts-of-speech tags to filter tokens, while summanlp's version uses a list of stopwords), and output only the extracted sentences.

I need some ways to bypass the text pre-processing function so I can use my own pipeline for different languages such as Chinese and Japanese. And I need to get the internal data structure to be able to visualize it.

After some investigations, I decided that summanlp/textrank was the easiest one for me to customize/extend (I really liked the abstraction of that project.).

The code used in this post has been open-sourced on Github, and comes with a simple web interface:

ceshine/textrank_demo

A simple website demonstrating TextRank's extractive summarization capability. - ceshine/textrank_demo

github.com

How does TextRank work?

We're going to focus on the sentence extraction part of the TextRank algorithm (i.e. ignoring the keyword extraction part). This answer on Quora did a great job explaining the intuition behind the algorithm, as quoted below:

> *What TextRank does is very simple:* ***it finds how similar each sentence is to all other sentences in the text. The most important sentence is the one that is most similar to all the others****, with this in mind the similarity function should be oriented to the semantic of the sentence, cosine similarity based on a bag of words approach can work well and BM25/BM25+ work really nicely for TextRank.*

Here are two good tutorials on TextRank algorithm: Document Summarization using TextRank, TextRank for Text Summarization. Please refer to them, or even better, read the paper[1] if you want to know more about TextRank.

An Example

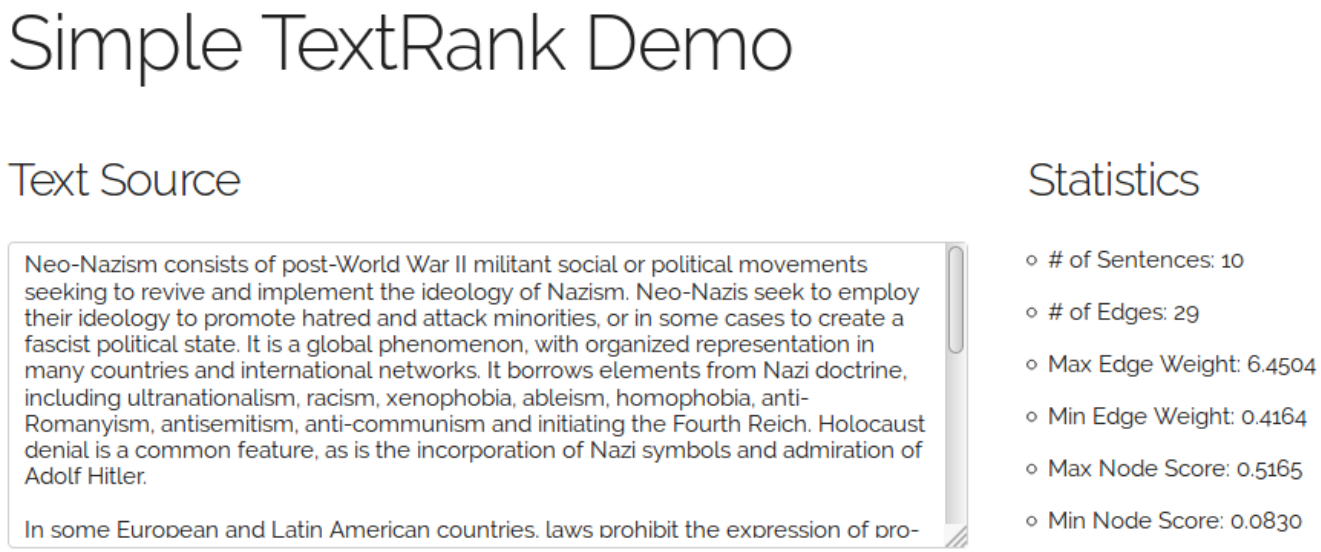
We're going to use the texts from the first two sections of the "Neo-Nazism" entry on Wikipedia as an example to demonstrate the web interface and to visualize the associated internal graph created by TextRank.

An static snapshot of the web interface can be found here:

Demo Page for TextRank Algorithm

Results Neo-Nazism consists of post-World War II militant social or political movements seeking to revive and implement...

publicb2.ceshine.net



The texts were split into 10 sentences, so there are 10 x 9 / 2 = 45 potential undirected connetctions between them. Only 29 of them has a weight larger than 0 (i.e. the sentences at both end of the edges are somewhat similar to each other). The similarity measure used here is the same as the original paper (something very similar to the Jaccard similarity coefficient):

Definition 1. *Given S_i, S_j two sentences represented by a set of n words that*

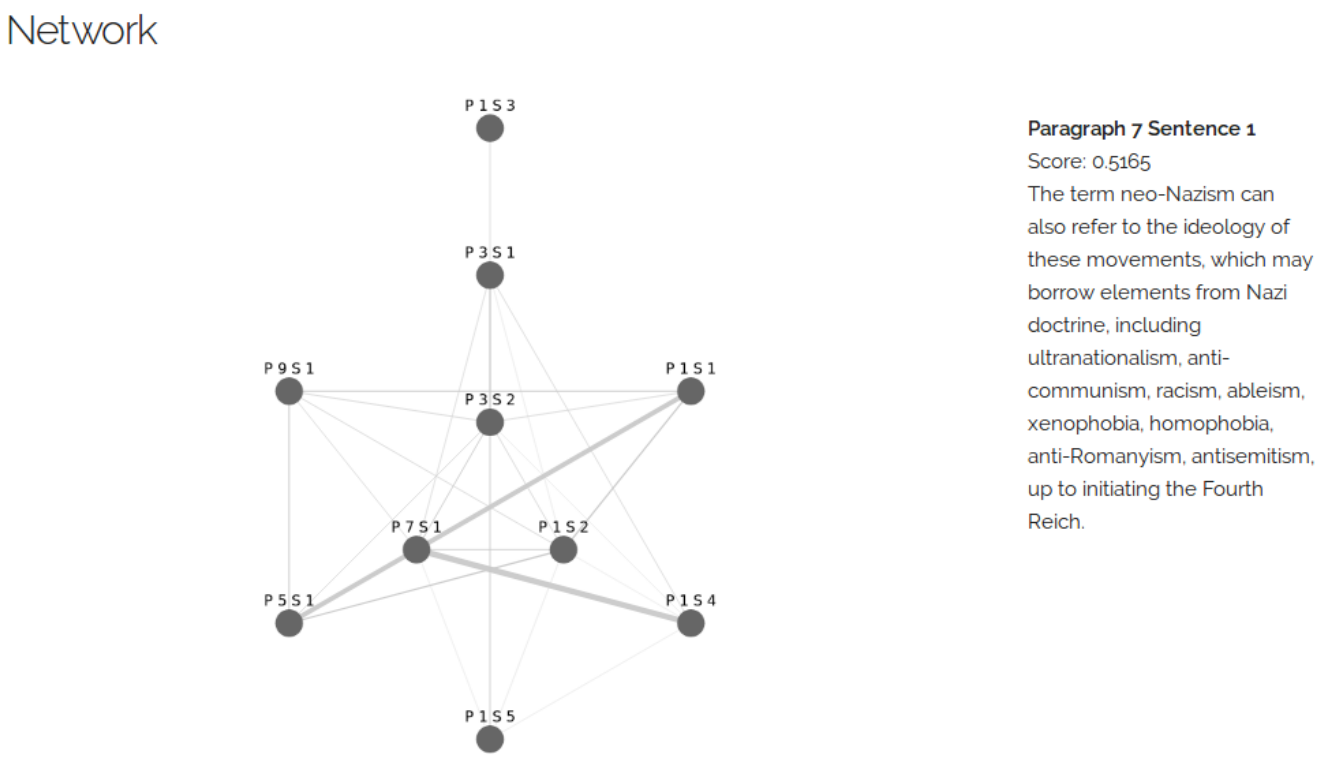
in S_i are represented as $S_i = w_1^i, w_2^i, ..., w_n^i$. The similarity function for S_i, S_j can be defined as:

$$Sim(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \quad (1)$$

Similarity function used. Source: [2]

However, the author summanlp/textrank provided several other measures in various branches of the repository, such as BM25+ with correction.

The Internal Graph



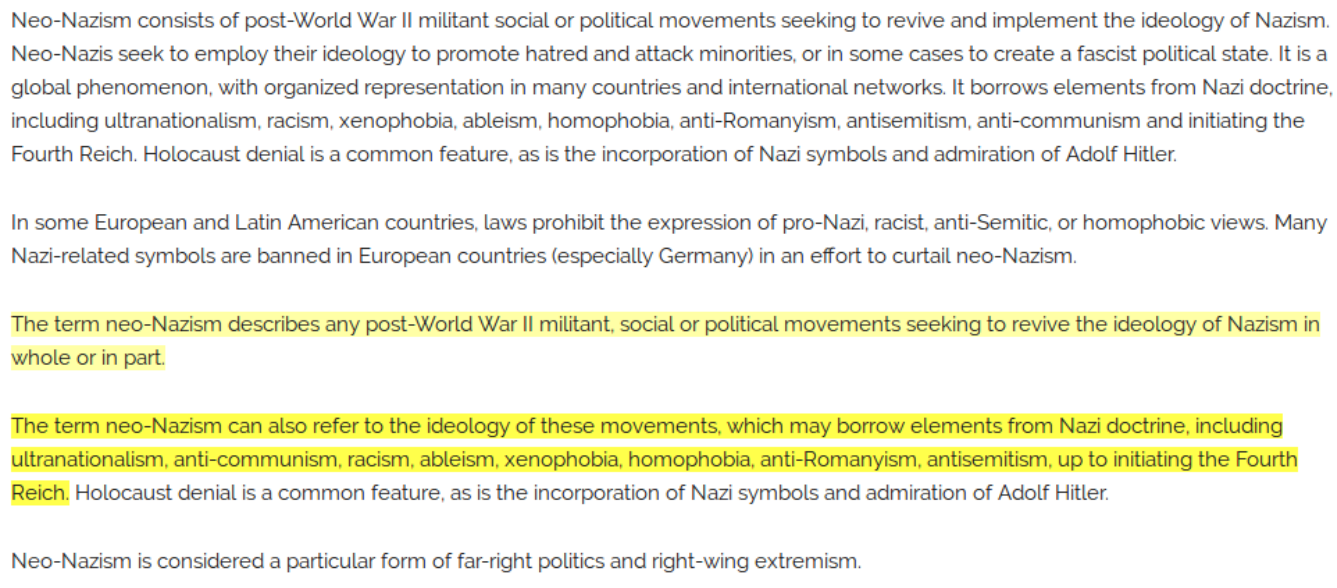
The graph constructed by TextRank (P1 S4 means the fourth sentence in the first paragraph)

The above plot actually has some limited interactivity. Please visit the static snapshot page, or clone and run the code from your machine to make use of it.

We can see that the first sentence of the seventh paragraph has some strong similarity with P1S1, P1S4 and P5S1 nodes. It implies that the ideas in that sentences had been (partly or fully) mentioned several times in the article, therefore it is most likely the centerpiece of the article.

(Note: the paragraph count includes empty lines.)

The Results



Neo-Nazism consists of post-World War II militant social or political movements seeking to revive and implement the ideology of Nazism. Neo-Nazis seek to employ their ideology to promote hatred and attack minorities, or in some cases to create a fascist political state. It is a global phenomenon, with organized representation in many countries and international networks. It borrows elements from Nazi doctrine, including ultranationalism, racism, xenophobia, ableism, homophobia, anti-Romanyism, antisemitism, anti-communism and initiating the Fourth Reich. Holocaust denial is a common feature, as is the incorporation of Nazi symbols and admiration of Adolf Hitler.

In some European and Latin American countries, laws prohibit the expression of pro-Nazi, racist, anti-Semitic, or homophobic views. Many Nazi-related symbols are banned in European countries (especially Germany) in an effort to curtail neo-Nazism.

The term neo-Nazism describes any post-World War II militant, social or political movements seeking to revive the ideology of Nazism in whole or in part.

The term neo-Nazism can also refer to the ideology of these movements, which may borrow elements from Nazi doctrine, including ultranationalism, anti-communism, racism, ableism, xenophobia, homophobia, anti-Romanyism, antisemitism, up to initiating the Fourth Reich. Holocaust denial is a common feature, as is the incorporation of Nazi symbols and admiration of Adolf Hitler.

Neo-Nazism is considered a particular form of far-right politics and right-wing extremism.

The texts with the most two important sentences highlighted

Looks reasonable enough. The highlighted sentences very broadly defines Neo-Nazism. Terms used in them frequently overlap the terms in other sentences.

Future Work

It appears that we already have a decent baseline for Wikipedia-like articles. However, I did not include a proper stop list in the program, so it might work terribly with more casual texts. Using POS tags could actually be a better idea as it is more robust than a fixed stop word list.

Bypassing the built-in text pre-processing procedure has not been tried yet. It is required for the program to be able to process other languages that is not supported by `SnowballStemmer`. We need to create `SyntacticUnit` objects from raw texts by ourselves.

The End

Thank you for reading! As TextRank is a fairly “old” algorithm by today’s standard, I wasn’t very sure how much efforts I need to put in to explain the algorithm. In the end I seem to have skipped most of the details.

The goal here is to provide a big picture and an example of how everything comes together for someone who is trying to quick build a baseline. I hope you find this post useful and worth your time (please give it some claps if you do). Thanks again for reading!

20181212 Update

I'm working on Chinese support, and have uploaded a very early version to the Github repo. You can check this file *text_cleaning_zh.py* for an example of a **custom pre-processing pipeline with a POS tag filter**. I'd probably also add Japanese support as well if I can find a good and free Japanese POS tag filter.

Results

这样的警告——科技巨头版的"大到不能倒"论调——有一种浅薄的民族主义号召力。毫无疑问，中国科技产业正在壮大，竞争力咄咄逼人，许多公司也都有政府的认可与支持。据统计，世界最大的20家科技企业中，有八家是中国公司。这似乎显示出了一种要争夺全球主导地位的态势，为此美国应该考虑的不是拆分或监管，而是尽一切可能地去保护和补贴自己的主队。

但接受这一论点将是错误的，因为它违背且忽视了一个用巨大代价换来的教训：以"民族骄傲"为中心的产业政策的荒唐之处，特别是在科技产业。真正想要的，是作为一个美国自己的社交媒体垄断企业得到认可与保护，在海外奋勇拼杀。但历史和基本的经济学常识都表明，我们应该相信国内的激烈竞争会使整个产业更强大，这样做的效果会好得多。

这是日美科技竞争的历史教训。在20世纪70年代和80年代初，人们普遍认为，日本正在威胁美国在科技市场的霸主地位。日本科技巨头NEC在主机市场上成为IBM的有力挑战者；索尼在消费电子产品领域突飞猛进，松下、东芝这些实力雄厚的公司也不甘示弱。日本政府通过通商产业省对这些企业进行支持，它奉行的民族主义产业政策被认为是万无一失的。

如果我们在反托拉斯执法上放这些企业一马，允许它们主导市场，并购竞争对手，美国可能会失去其标志性的优势：允许新旧替换、接受反叛和改变的意愿——托马斯·杰弗逊(Thomas Jefferson)的反叛循环和"爱国者的鲜血"的工业版。

那么，正如扎克伯格所预言的，科技的未来可能真的就要属于中国了。

Network

Paragraph 1 Sentence 2

Score: 0.4071

毫无疑问，中国科技产业正在壮大，竞争力咄咄逼人，许多公司也都有政府的认可与支持。

Tokens: 中国 科技 产业 壮大 竞争 力 公司 有 政府 认可 支持

Built by CeShine Lee, based on TextRank implementation summanlp/textrank.

A snapshot of the Chinese support (WIP)

References:

1. Mihalcea, R., Tarau, P.: “Textrank: Bringing order into texts”. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004. pp. 404–411. Association for Computational Linguistics, Barcelona, Spain. July 2004.

2. Barrios, F., López, F., Argerich, L., Wachenchauzer, R.: “Variations of the Similarity Function of TextRank for Automated Summarization”. Anales de las 44JAIIO. Jornadas Argentinas de Informática, Argentine Symposium on Artificial Intelligence, 2015.

(This post is also published on my personal blog.)

Machine Learning NLP Visualization Data Science