Project Report  
Search Engines and Information Retrieval Systems

Project 8: Wikipedia Information – Group 12

Ambjörn Karlsson  
Daniel Mareno Romuld  
Axel Samuelsson  
Charles-Elie Simon  
Jian Sun

# Abstract

Abstract goes here.

# Introduction

## TEXT SUMMARIZATION

Wikipedia contains a wealth of information, sometimes too much so. That makes it a fantastic resource for researching a subject thoroughly, but can also make it difficult to learn more about a subject at a glance.

Take, for example, the article about *Toilet paper orientation* [1]. Excluding references, printing it yields a 13 page PDF including everything from scientific surveys, breakdowns by age to analysis of arguments, et cetera. However, the base subject is not really particularly complex, and quite easily summed up by a human: “The topic of toilet paper roll orientation is surprisingly divisive. The roll can be in either the ‘over’ (the end hanging away from the wall) or ‘under’ (the end hanging towards the wall) orientation, and a lot of people have very strong opinions about which is better.”

The goal of this project is to get a computer to perform the same task. Give the program a Wikipedia article, or just a subject / search string, and it returns a summary of the given topic.

## FOI

The customer for this project is The Swedish Defence Research Agency - FOI. FOI is a research institute run by the Swedish government, tasked with coordinating research between military and civilian research efforts, in order to improve Sweden’s research efforts from a holistic perspective.

# Background

## AUTOMATIC SUMMARIZATION

There are two main types of automatic summarization: extraction-based (EBS) and abstraction-based (ABS). EBS aims to extract objects (generally sentences, but words are possible if the aim is tagging rather than a natural language summary) from the original text and use them as a summary without modification. ABS is, as the name hints, more abstract. Instead of extracting whole pieces and puzzling them together, ABS aims to take an approach more similar to that of a human, paraphrasing sections and using information from the original text rather than the text itself.

For this project we will be performing EBS, comparing a few different algorithms to one another as well as reference summaries written by ourselves.

## EXTRACTION-BASED SUMMARIZATION ALGORITHMS

There are a large number of approaches for EBS, but we’ve chosen to narrow our scope to the two most popular algorithms.

First, there is the naïve solution, used by among others a reddit bot that creates summaries of news articles. It’s simple. Pick the first sentence of every paragraph, as well as the last sentence of the last paragraph. The thought behind it is that sentences that begin a paragraph are likely to work well on their own, grammatically, as there is no previous sentence in the paragraph to refer back to. The last sentence is chosen because news articles are often written in such a way that the last sentence wraps everything up. This approach may not translate perfectly for Wikipedia, especially not with the summarizing last sentence, but still shows some promise.

Then, we have what appears to be the most popular option, LexRank. It is an unsupervised machine learning technique, which means it doesn’t require pre-tagged training and testing data. LexRank is very similar to TextRank, which was developed simultaneously by a separate team, for purposes of tagging texts rather than summarizing them. They are both, in turn, heavily inspired but the PageRank algorithm for ranking search results. Each sentence is represented as a vertex, then LexRank uses cosine similarity of the TF-IDF-vectors of sentences to create edges between sentences, and then performs a random walk over the graph to find the most “central” sentence(s) of the document.

## ROUGE

The ROUGE metric is used to compare a machine summary to a reference (human) summary[2]. For this project we’ve decided to use the 1-gram evaluation metric (ROUGE-1). It’s easy to grasp because of its simplicity, and relatively simple to implement, giving us more time to focus on the summarization itself. ROUGE-1 returns a score between 0 and 1 – the fraction of the words from the reference summary that are present in the machine summary.

# Method

Method goes here.

# Results

Results go here.

# Discussion

Discussion goes here.

# References

1. <https://en.wikipedia.org/wiki/Toilet_paper_orientation>, accessed 2016-04-18 11:31.
2. <http://www.aclweb.org/anthology/W/W04/W04-1013.pdf>, Chin-Yew Lin, 2004.