Machine Learning in Keyword Extraction

One cool trick ensures you never have to read a full paper again!

Authors HATE this simple trick!

Researcher: Aditya Subramanian
Mentor: Prof. Tom Kaczmarek

What is Keyword Extraction?

The process by which a program automatically identifies terms that represent key concepts in a text document.

Why is it useful?

- It helps search algorithms find and filter through relevant articles
- It helps you not read long dry papers and articles

How do we do it?

There are many techniques, including:

- Find term frequency, Inverse Document frequency, and Part of speech
- Recursive graph based approach (not used here)

Acknowledgements

Based on research by Hulth (2003) Improved Automatic Keyword Extraction Given More Linguistic Knowledge
Corpus used was from Krapivin (2009) Exploring and Understanding Citation-based Scientific Metrics
Data extraction was largely possible due to Stanford’s Natural Language Tool Kit
This work was supported in part by a National Science Foundation REU Site grant, #ACI-1461264, 'Computation Across the Disciplines', at Marquette University.

What did we do?

I proposed new factors for extracting keywords. They will then be tested in TF-IDF and graph based algorithms to compare them to current standards.

First I compiled data for these factors from a 1000 document corpus. I then ran a logistic regression to appropriately weight each factor. I selected the best model based on fit and interpreted the results.

What did we test?

- *Term Frequency (TF):* How many times a term is in the text
- *Inverse Document Frequency (IDF):* How many other documents in the corpus did the term appear in?
- *Part of Speech (PoS):* Is the term a verb? Noun? Adjective? Pronoun?
- *First/Last word in sentence:* How often is the term the first or last word in a sentence it appears in?
- *In Intro/Conclusion:* Does the term appear in the introduction or conclusion of the paper?
- *Most Frequent Ngram:* How many times does the same 2 or 3 term phrase containing this word appear?

What did we find?

Increase keyword likelihood	Decrease keyword likelihood
Nouns, Plural Proper Nouns, Verbs, Adjectives, and Adverbs	Plural Nouns, Pronouns, Conjunctions, and other types of words
Term Frequency and TF*IDF	Inverse Document frequency
Last word in Sentence	(Log) Length of sentence
In Intro, in conclusion	
In both intro AND conclusion	
Biggest Ngram	

Next Step

The next step is to see how well the algorithm predicts keywords using the new factors, and then compare that to other similar algorithms.

Aditya Subramanian | Carleton College ’17 | Computer Science and Economics | subramaniana@carleton.edu