Categorization of research papers

IRE Major Project Phase 2

- Team 54 (201301203, 201330072, 201505544)

Problem Statement:
Assign a category to given research paper from a list of predefined categories as mentioned in 2012 ACM Computing Classification System .

The Test Documents Considered:
We have considered abstract of ACM research papers as a document and it'l be annotated to a closely related category in the 2012 ACM Computing Classification System. Accuracy is calculated based on whether the document is classified to the category it actually belongs to.

Training Data:
Summary of wikidata of each of the leaf categories in the ACM categories is taken for training the model.

Testing Data:
Abstract of a research paper (ACM)

Models Used:

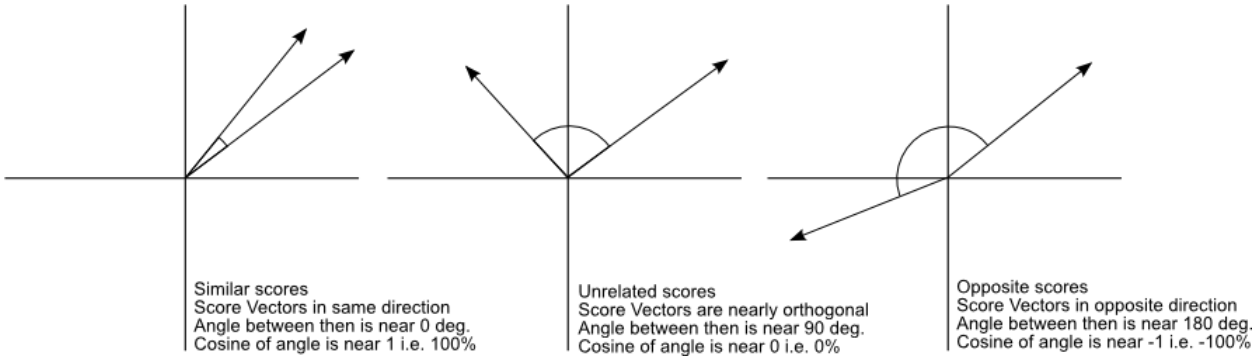
1. **The Cosine Similarity**

In our approach we took the wikipedia summary of each of the categories as separate documents and measured the cosine similarity of the test document with each of them and assigned the category whose wikipedia document was most similar.
Brief overview of cosine similarity :
The cosine similarity between two documents on the Vector Space is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents. What we have to do to build the cosine similarity equation is to solve the equation of the dot product for the $\cos\theta$:

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

And that is it, this is the cosine similarity formula. Cosine Similarity will generate a metric that says how related are two documents by looking at the angle instead of magnitude, like in the examples below:



The Cosine Similarity values for different documents, 1 (same direction), 0 (90 deg.), -1 (opposite directions).

Note that even if we had a vector pointing to a point far from another vector, they still could have an small angle and that is the central point on the use of Cosine Similarity, the measurement tends to ignore the higher term count on documents. Suppose we have a document with the word “sky” appearing 200 times and another document with the word “sky” appearing 50, the Euclidean distance between them will be higher but the angle will still be small because they are pointing to the same direction, which is what matters when we are comparing documents.

Reason for choosing this :

Since the categories will be similar to the abstract of the research paper, we implemented this. We will try to expand this using doc2vec to even take care of semantic similarity in the next deliverable.

- Latent Dirichlet Allocation

In our approach, we used LDA (topic modelling algorithm) to cluster documents and assign a category to each cluster depending on the cluster in which the wikipedia document of the category is matching the most. Then put the test document on same model and assigned the category depending on the cluster to which test document belonged and the category assigned to that cluster.

Brief overview of LDA:

In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the pLSA model is equivalent to the LDA model under a uniform Dirichlet prior distribution.[5]

For example, an LDA model might have topics that can be classified as **CAT_related** and **DOG_related**. A topic has probabilities of generating various words, such as *milk*, *meow*, and *kitten*, which can be classified and interpreted by the viewer as "CAT_related". Naturally, the word *cat* itself will have high probability given this topic. The **DOG_related** topic likewise has probabilities of generating each word: *puppy*, *bark*, and *bone* might have high probability. Words without special relevance, such as *the* (see function word), will have roughly even probability between classes (or can be placed into a separate category). A topic is not strongly defined, neither semantically nor epistemologically. It is identified on the basis of supervised labeling and

(manual) pruning on the basis of their likelihood of co-occurrence. A lexical word may occur in several topics with a different probability, however, with a different typical set of neighboring words in each topic.
Each document is assumed to be characterized by a particular set of topics. This is akin to the standard bag of words model assumption, and makes the individual words exchangeable.

Results:

Cosine Similarity:

Test Data #	Category Accuracy	Tag Accuracy
10	30%	80%

Note: Cases where category is missing, tag accuracy is considered to calculate category accuracy

Paper Name	Actual Category	Result
The You in Youtube	Human Information Processing	Youtube (correct)
Enhanced Skype Traffic Identification	None	Skype (correct)
Take a close look at phishing	Internet	Phishing (correct)
Realtime java technology in avionics	None	Software Evolution (wrong)
MapReduce Algorithms	None	Parallel Algorithms (correct)
Collaborative Energy Conservation in a microgrid	Special-Purpose and Application-based systems	Power and Energy (correct)
Proposal of privacy protection system for web forms using bloom filter	Security and Protection	Social Networking Sites (wrong)
Spoilt for Choice - Graph-Based Assessment of of key management protocols	Data Sharing	Data Mining (correct)

to share encrypted data		
Unsolved problems in search	Information Storage and Retrieval	Enterprise search (correct)
Correlation Clustering	None	Cluster Analysis (correct)

Latent Dirichlet Allocation:

Test Data #	**Category Accuracy**	**Tag Accuracy**
10	30%	60%

Note: Cases where category is missing, tag accuracy is considered to calculate category accuracy

Paper Name	**Actual Category**	**Result**
The You in Youtube	Human Information Processing	Youtube (correct)
Enhanced Skype Traffic Identification	None	Biometrics (correct)
Take a close look at phishing	Internet	III-V Compounds (wrong)
Realtime java technology in avionics	None	ANSI C (wrong)
MapReduce Algorithms	None	Parallel Programming (correct)
Collaborative Energy Conservation in a microgrid	Special-Purpose and Application-based systems	Formal Software Verification(wrong)
Proposal of privacy protection system for web forms using bloom filter	Security and Protection	Collaborative Filtering (wrong)
Spoilt for Choice - Graph-Based Assessment of of key management protocols	Data Sharing	Public Key Encryption(correct)

to share encrypted data		
Unsolved problems in search	Information Storage and Retrieval	Web Search Engines (correct)
Correlation Clustering	None	Cluster Analysis (correct)

Challenges:

- Since the size of training data is 1636 each for a leaf category in ACM classification, and if a clustering technique like LDA is used, there'll be 1636 clusters. But each cluster will not have enough data to semantically match and classify a test document (abstract) to its correct category.

- If we use a similarity based model like “cosine similarity”, the document is matches to related but wrong category. This is because cosine similarity doesn’t take semantics into consideration. For ex, let’s consider the abstract

“In this talk, I will argue that complex sociotechnical systems like YouTube admit multiple conceptualizations - what YouTube is and is about - and that each of these perspectives ultimately results in different ways of posing research questions in multimedia. I will then contrast the existing lines of YouTube multimedia research under this view. In particular, I will discuss a perspective that put the focus on people - the You in YouTube – and present an overview of ongoing work that aims at developing computational models to characterize YouTube as a collection of communities where individuals express and communicate through video. I will finally discuss opportunities for future research in multimedia under this framework.”

- Because of the presence of “youtube” many times, the document is classified to category “YouTube” instead of “Human Information Processing” to which the document actually belongs to.

Future Work:
Use a model to semantically classify paper abstracts using the limited wiki data available.

Test Data Papers:

1. http://dl.acm.org/citation.cfm?id=1878153&CFID=596261973&CFTOKEN=22019760
2. http://dl.acm.org/citation.cfm?id=1345297&CFID=596261973&CFTOKEN=22019760
3. http://dl.acm.org/citation.cfm?id=1409943&CFID=596261973&CFTOKEN=22019760
4. http://dl.acm.org/citation.cfm?id=1850791&CFID=596261973&CFTOKEN=22019760
5. http://dl.acm.org/citation.cfm?id=2778866&CFID=596261973&CFTOKEN=2201976 0

6. http://dl.acm.org/citation.cfm?id=2674079&CFID=596261973&CFTOKEN=22019760
7. http://dl.acm.org/citation.cfm?id=2184765&CFID=596261973&CFTOKEN=22019760
8. http://dl.acm.org/citation.cfm?id=2674079&CFID=596261973&CFTOKEN=22019760
9. http://dl.acm.org/citation.cfm?id=1458085&CFID=596261973&CFTOKEN=22019760
10. http://dl.acm.org/citation.cfm?id=2630808&CFID=596261973&CFTOKEN=22019760