2. Read the following excerpts and decide to which section of their research articles they belong: introduction, methods, results or conclusion. What helped you decide? Discuss in class.

a. The physicochemical properties of biodiesel have attracted much attention in the last years. As a mixture of different alkyl esters, fatty acid methyl esters (FAMEs) or fatty acid ethyl esters (FAEEs), biodiesel composition depends on the raw materials that are used. Vegetal oils and some kinds of animal tallows are typically used [1] to produce biodiesel but depending on the composition of the used raw materials, biodiesel properties can show a low quality performance in some aspects.[23]

b. Two density models (theoretical and empirical) were evaluated for methanol and ethanol based biodiesels with a large number of experimental and bibliographical data. These models were also improved using the new reported data rendering a good final accuracy for the prediction of biodiesel density at 15 _C. In the case of methanol based biodiesels, the evaluated and improved methods can be considered suitable for using it in a biodiesel simulation or in a biodiesel production plant.[24]

Automatic Classification of Computer Generated Papers

Allen Lavoie – Rensselaer Center for Open Software (RCOS)

Purpose

Identify computer generated academic papers (e.g. MIT's SCIgen).

Methods

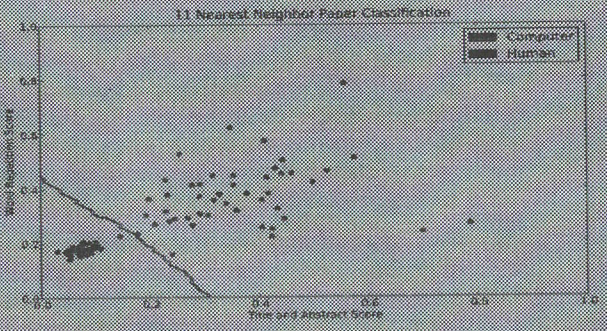
1. Pre-processing: Convert to text, tokenize, filter parts of speech and stem
2. Feature based scoring
   (a) Occurrences of words from the title and abstract in the paper's body
   (b) Occurrences of top ten most used words vs. all other words
3. Classification: Nearest neighbor (kd-tree)

Results

98% of papers classified correctly (100 samples)

'peertop', commun', 'internet', 'commun', 'share', 'resourc', 'entiti', 'end', 'user', 'compul', 'client', 'server', 'clientserv', 'paradigm' 'focu', 'peertop', 'network', 'type', 'bittorr', 'mean', 'speci', 'network', 'topolog', 'particip', 'peer', 'network', 'replic', 'constant', 'network', 'exchang', 'piec', 'list', 'bittorr', 'work', 'consid', 'replic', 'arriv', 'peer'

Text after pre-processing



Future work

- Web service for classification (Google App Engine)
- Investigate additional features
- Improve pre-processing efficiency
- Challenge SCIgen maintainers to avoid detection

Paper Sources

Computer: SCIgen http://pdos.csail.mit.edu/scigen/

Human: arXiv http://arxiv.org/

Key Technologies

- Python
- Natural Language Toolkit (NLTK)
- Scipy, Pylab, Matplotlib
- PDFMiner

Further Reading

Blog: http://paperdetection.blogspot.com/

Code: http://code.google.com/p/paper-detection/

RCOS: http://rcos.rpi.edu