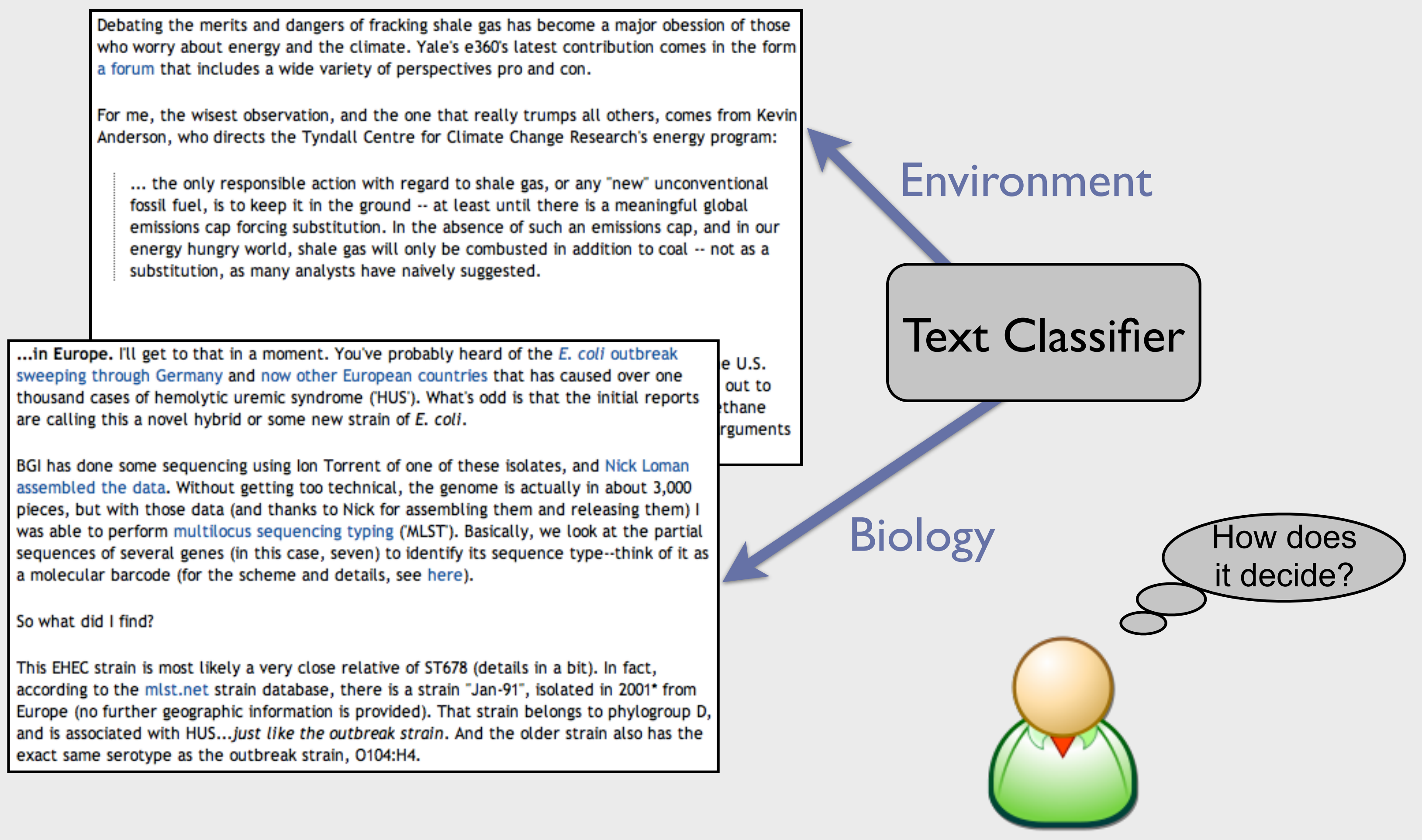
Visualizing Text Classification Models with Voronoi Word Clouds

Christin Seifert, Wolfgang Kienreich und Michael Granitzer
University of Technology & Know-Center, Graz, Austria
{cseifert, wkien, mgrani}@know-center.at

Motivation



Debating the merits and dangers of fracking shale gas has become a major obession of those who worry about energy and the climate. Yale's e360's latest contribution comes in the form a forum that includes a wide variety of perspectives pro and con.

For me, the wisest observation, and the one that really trumps all others, comes from Kevin Anderson, who directs the Tyndall Centre for Climate Change Research's energy program:

> ... the only responsible action with regard to shale gas, or any "new" unconventional fossil fuel, is to keep it in the ground -- at least until there is a meaningful global emissions cap forcing substitution. In the absence of such an emissions cap, and in our energy hungry world, shale gas will only be combusted in addition to coal -- not as a substitution, as many analysts have naively suggested.

e U.S.
out to
thane
rguments

...in Europe. I'll get to that in a moment. You've probably heard of the *E. coli* outbreak sweeping through Germany and now other European countries that has caused over one thousand cases of hemolytic uremic syndrome ('HUS'). What's odd is that the initial reports are calling this a novel hybrid or some new strain of *E. coli*.

BGI has done some sequencing using Ion Torrent of one of these isolates, and Nick Loman assembled the data. Without getting too technical, the genome is actually in about 3,000 pieces, but with those data (and thanks to Nick for assembling them and releasing them) I was able to perform multilocus sequencing typing ('MLST'). Basically, we look at the partial sequences of several genes (in this case, seven) to identify its sequence type--think of it as a molecular barcode (for the scheme and details, see here).

So what did I find?

This EHEC strain is most likely a very close relative of ST678 (details in a bit). In fact, according to the mlst.net strain database, there is a strain "Jan-91", isolated in 2001* from Europe (no further geographic information is provided). That strain belongs to phylogroup D, and is associated with HUS...*just like the outbreak strain*. And the older strain also has the exact same serotype as the outbreak strain, O104:H4.

Text Classification

Text Preprocessing

- tokenization, stemming, stop-words, POS-Tags
- vector-space model, nouns + proper nouns

openNLP

Class-Feature Centroid Classifier - CFC [1]

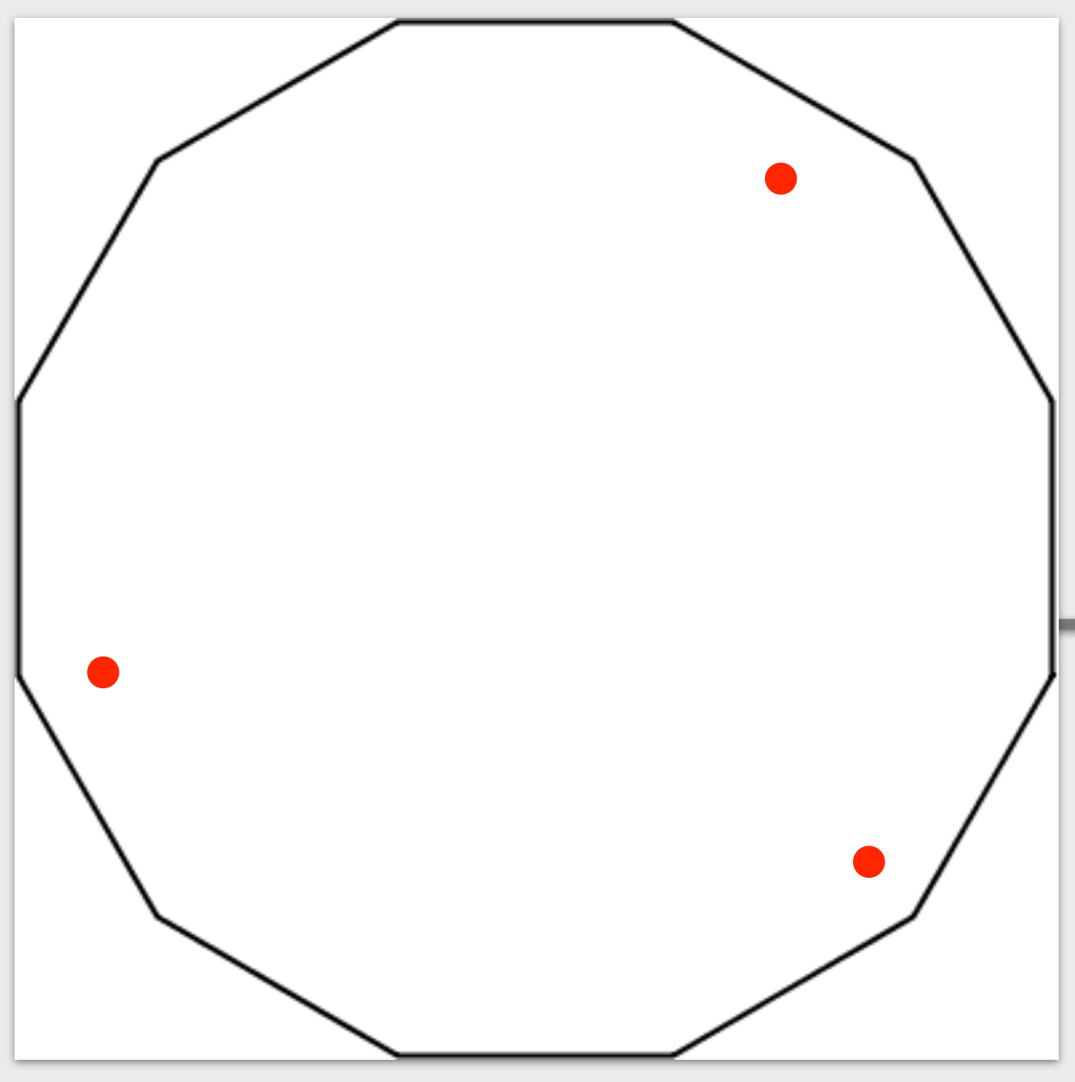
- centroid-based classification
- weighting schema for centroids designed specifically for text classification
- weight of a term in the centroid vector determined by inner-class term index and inter-class term index

$$w_{ij} = b^{\frac{DF^j_{t_i}}{|C_j|}} \times \log(\frac{|C|}{CF_{t_i}})$$

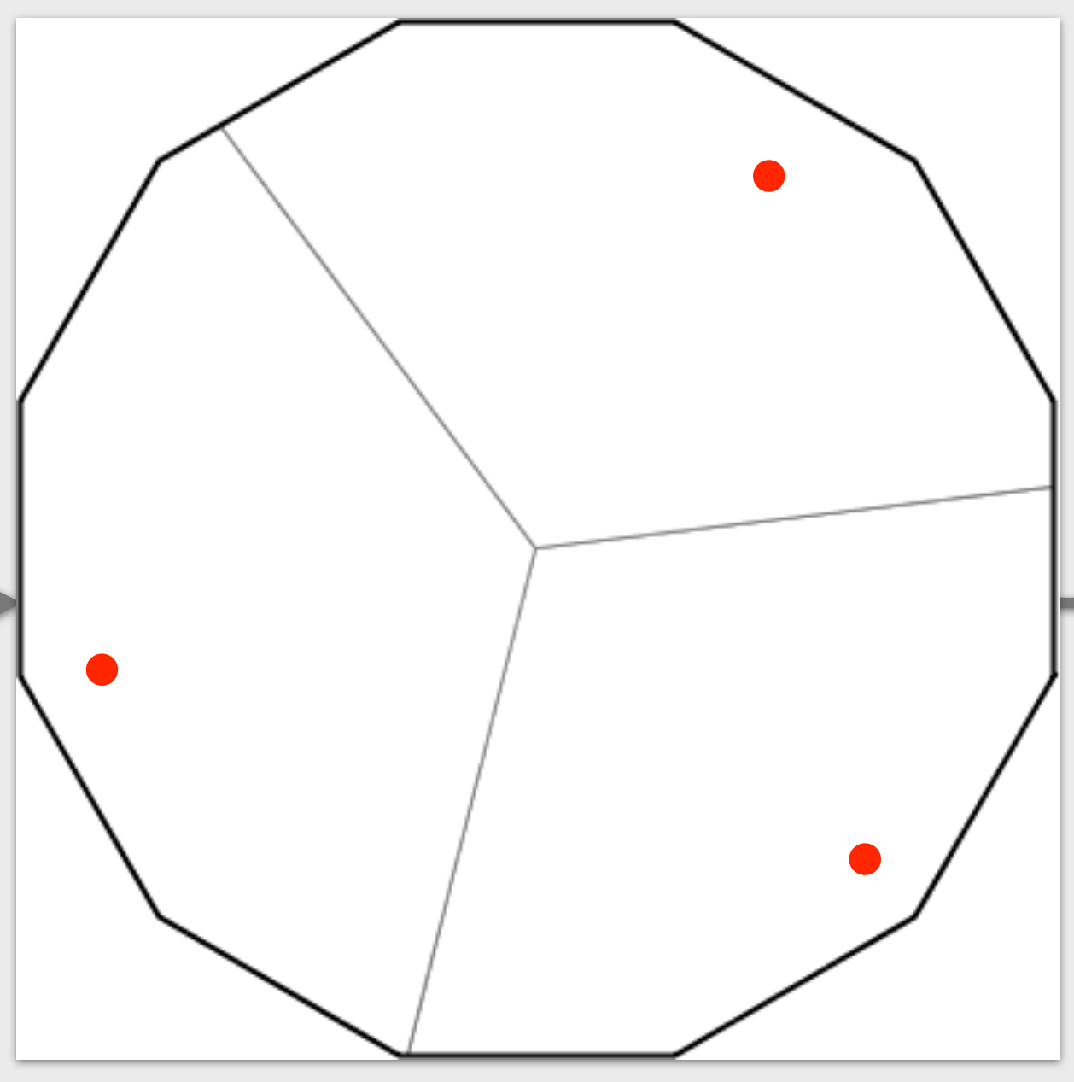
Output

- feature vector in noun-vector space for each class (=centroid)
- dictionary lookup of terms with highest weight in feature vector to get most important terms for each class

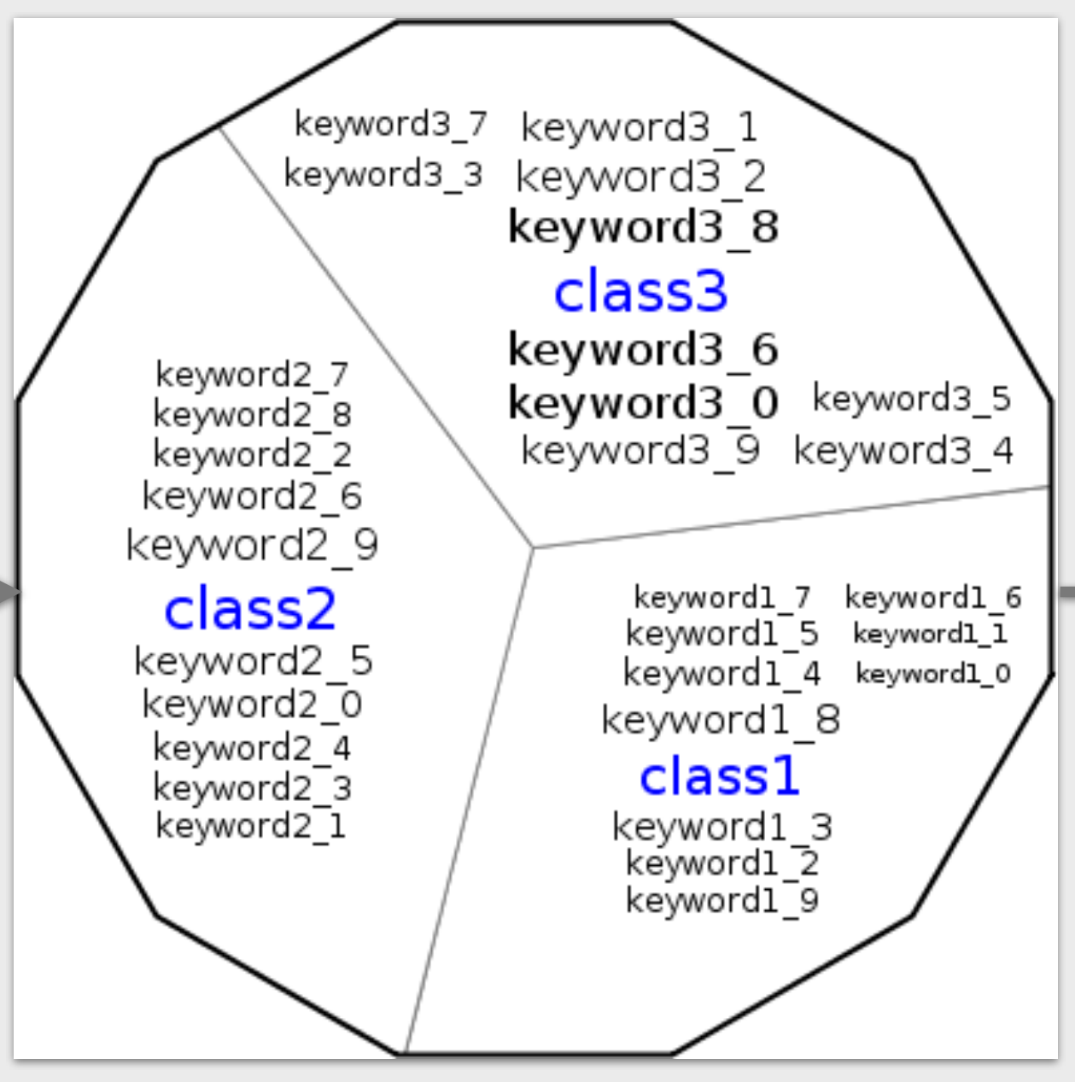
Constructing the Visualization



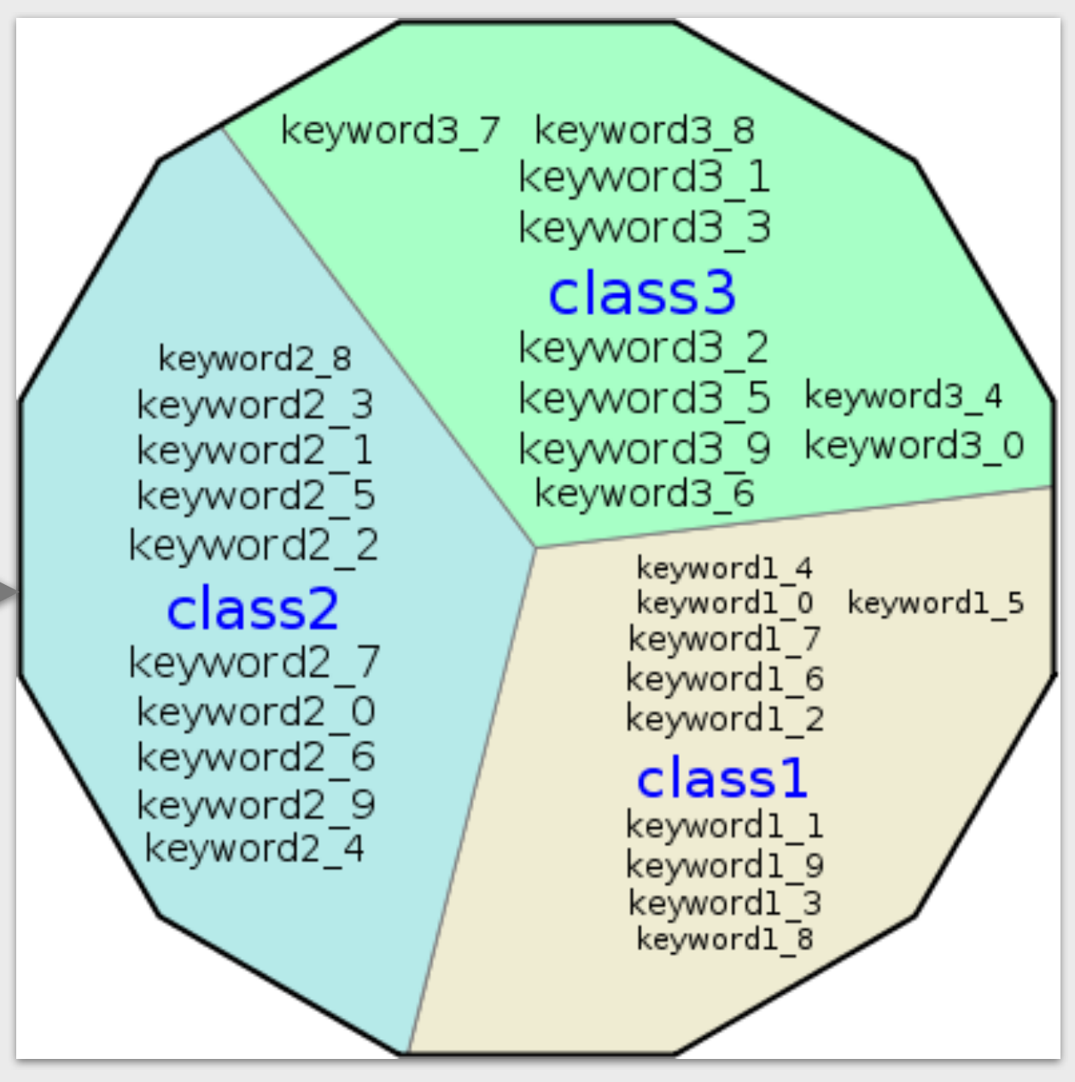
similarity layout of of class centroid vectors



class centroid vectors as generator points for Voronoi diagram



layout of class keywords in each Voronoi region using layout algorithm from [2]



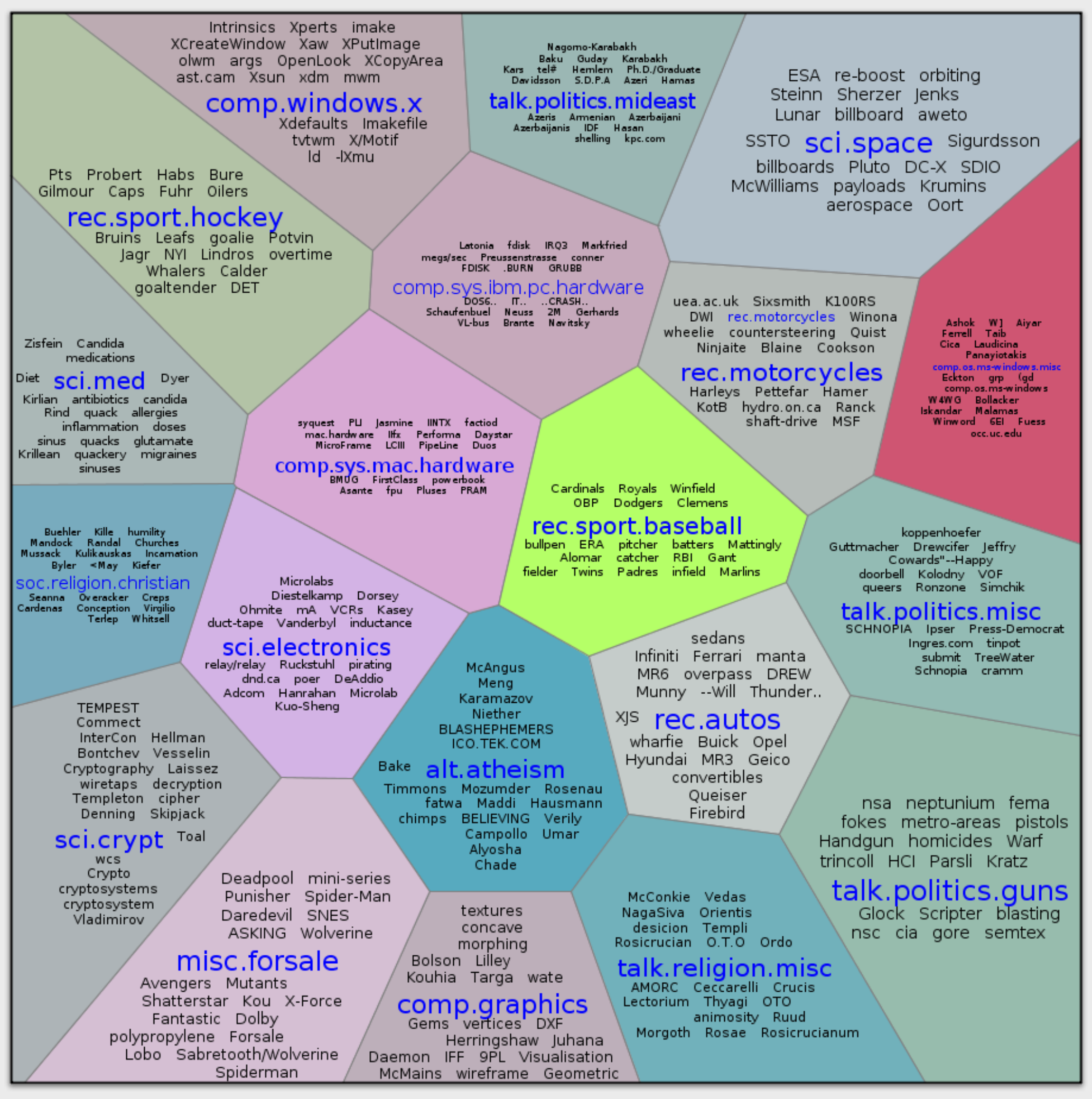
coloring of regions according to class similarity

Coloring

- FDP of class centroids to 3D-space
- RGB color for Voronoi region determined by

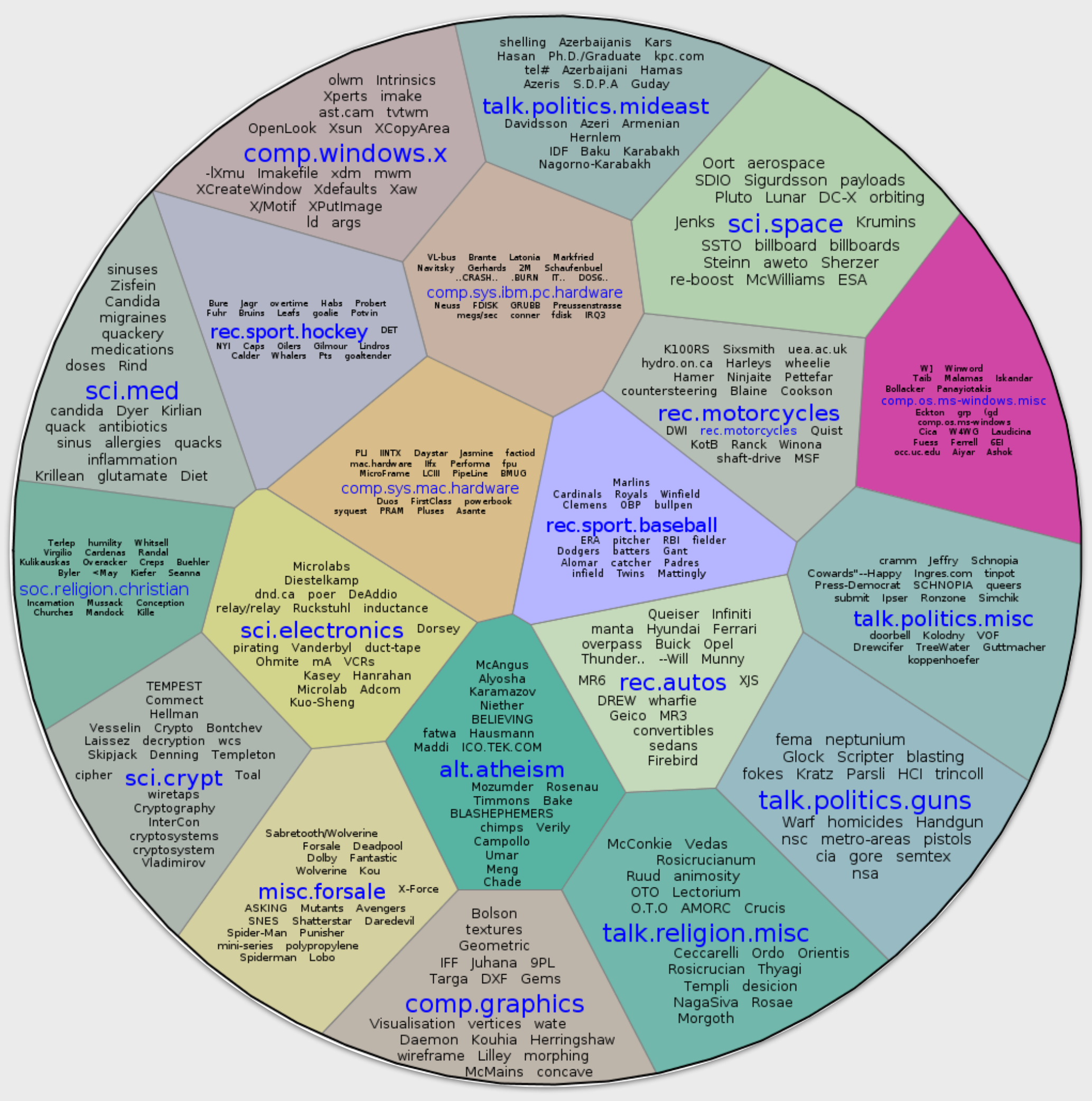
r = min(255, 255 · (0.7 + 0.7x))
g = min(255, 255 · (0.7 + 0.7y))
b = min(255, 255 · (0.7 + 0.7z))

Example - 20 Newsgroup Data Set



Top 10 terms

1.article
2.people
3.X
4.time
5.way
6.God
7.system
8.anyone
9.something
10.problem



Future Directions

- *improve coloring*
  similar classes should have similarly perceived colors, but still be distinguishable
- *interactive visualization*
  zoom, number of keywords vary with LoD
- *interactive machine learning*
  user can add or remove keywords for classes and the internal model of the classifier is updated

References

[1] Hu Guan, Jingyu Zhou, and Minyi Guo. A class-feature- centroid classifier for text categorization. In Proc. of the Inter- national conference on World Wide Web (WWW), pages 201– 210, New York, NY, USA, 2009. ACM.
[2] Christin Seifert, Barbara Kump, Wolfgang Kienreich, Gisela Granitzer, and Michael Granitzer. On the beauty and usability of tag clouds. In Proceedings of the 12th International Conference on Information Visualisation (IV), pages 17–25, Los Alamitos, CA, USA, July 2008. IEEE Computer Society.

Christin Seifert
Knowledge Management Institute, Graz University of Technology
Know-Center
cseifert@know-center.at

KNOW Center
TU Graz

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies.
Under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor
and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

COMET
Competence Centers for Excellent Technologies