T-61.271 Information visualization

Exercise 7. Wed 24.11.2004:10-12 T3

- Focus and context
- Visualizing text

1. Fisheye view.

   a. Assume that a data structure (e.g. contents of a book) that can be represented by a tree of height m and brancing factor b. How many steps does it take to traverse from one leaf node to another by traversing though fisheye-views of the tree?

   b. Present a (short) program code of you choice by setting the focus on some line and using the fisheye view.

2. The size of vocabulary can be reduced by removing some rare and very common words ("and", "the", ...). What could you do to reduce the vocabulary further?

3. How are the distances and the resulting document presentations in 2D space affected if the document vectors are not normalized (not normalized means that $|d_a|^2$ is generally not equal to $|d_b|^2$, $a \neq b$)?

4. Random projection can be used to reduce the dimension of the document vectors: $d_a \to d_a R$, Under what condition does the *expectation* of the similarity between two document vectors ($SIM(a,b) = d_a d_b^T$) remain unchanged (i.e. how should you generate entries for the matrix R)? How do you expect the *variance* of the similarity $SIM(a,b)$ behave as a function of the final dimensionality M' of the document vector?

   [Hints:
   `http://www.hut.fi/Yksikot/Kirjasto/Diss/2000/isbn9512252600/`, article 4,
   `http://www.cis.hut.fi/sami/abstracts.html#ijcnn98`]

5. Dimension reduction methods, such as the use of feature words, random projection or latent semantic indexing can be useful if you use PCA or SOM to present the documents in 2D space. However, these dimension reduction methods are useless if you use MDS. Why? Would you expect to be able to represent the one million documents of WEBSOM by using MDS instead of SOM?