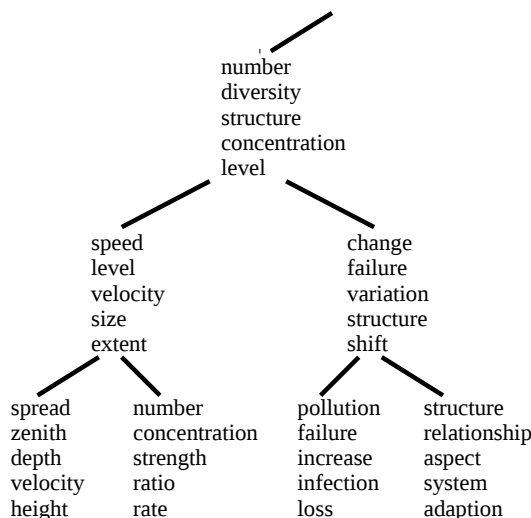


I. From Table 1 of Hofmann (1999): Four learned topic models ($k = 128$) drawn from the TDT-1 corpus (15,862 documents of newswire text and transcripts), each represented by the words with highest probability with respect to that model.

- 1 plane, airport, crash, flight, safety, aircraft, air, passenger, ...
- 2 space, shuttle, mission, astronauts, launch, station, crew, ...
- 3 home, family, like, love, kids, mother, life, happy, friends, ...
- 4 film, movie, music, new, best, hollywood, love, actor, entertainment, ...

II. From Figure 2 of Pereira, Tishby, and Lee (1993): A portion of a cluster hierarchy learned from Grolier's Encyclopedia (ten million words). Each node in the "tree" below represents a cluster by listing the five words most closely associated with that cluster. Cluster splits are represented by parent-child relationships in the tree.



Note: In these experiments, the 1000 most common nouns played the role of documents, and verbs that take those nouns as direct objects played the role of terms. Thus, the noun "gun" could be characterized by its appearing in the verb/direct-object relationship with the verbs "fire", "polish", "buy", and "ban"; whereas the noun "banana" could be characterized by its appearing in the verb/direct-object relationship with the verbs "peel", "eat", "buy", and, conceivably, "ban".

III. Evaluation note Inferred structure(s) should be shown to provide improvement in (more or less) practical applications, rather than only "eyeballed" for correspondence to intuitions regarding language semantics. Recall, after all, that the singular vectors may not be "interpretable", but SVD-based subspace projections often are very effective.

References

- Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR*, pages 50–57.
- Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *31st Annual Meeting of the ACL*, pages 183–190. ACL, Association for Computational Linguistics, Somerset, NJ. Distributed by Morgan Kaufmann, San Francisco, CA.