A simple document model:

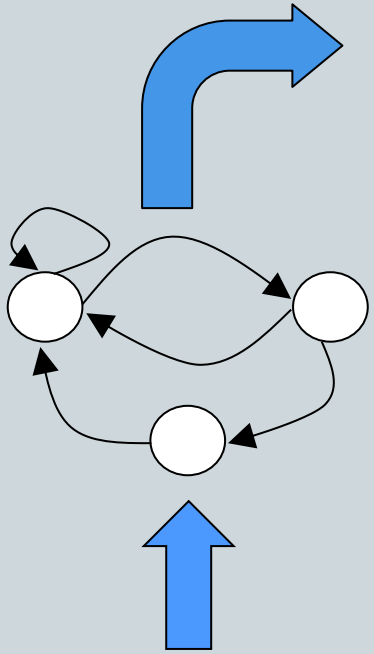
Bag-of-Words

KING'S College LONDON

NIHR | Maudsley Biomedical Research Centre

Rationalism: pattern matching

- Rationalist, armchair approach
- Brittle
- Not bad for simple regular patterns
- Hard to scale



```
/[0-9]+ *[Mm].?[Gg].?/
```

The rise of empiricism

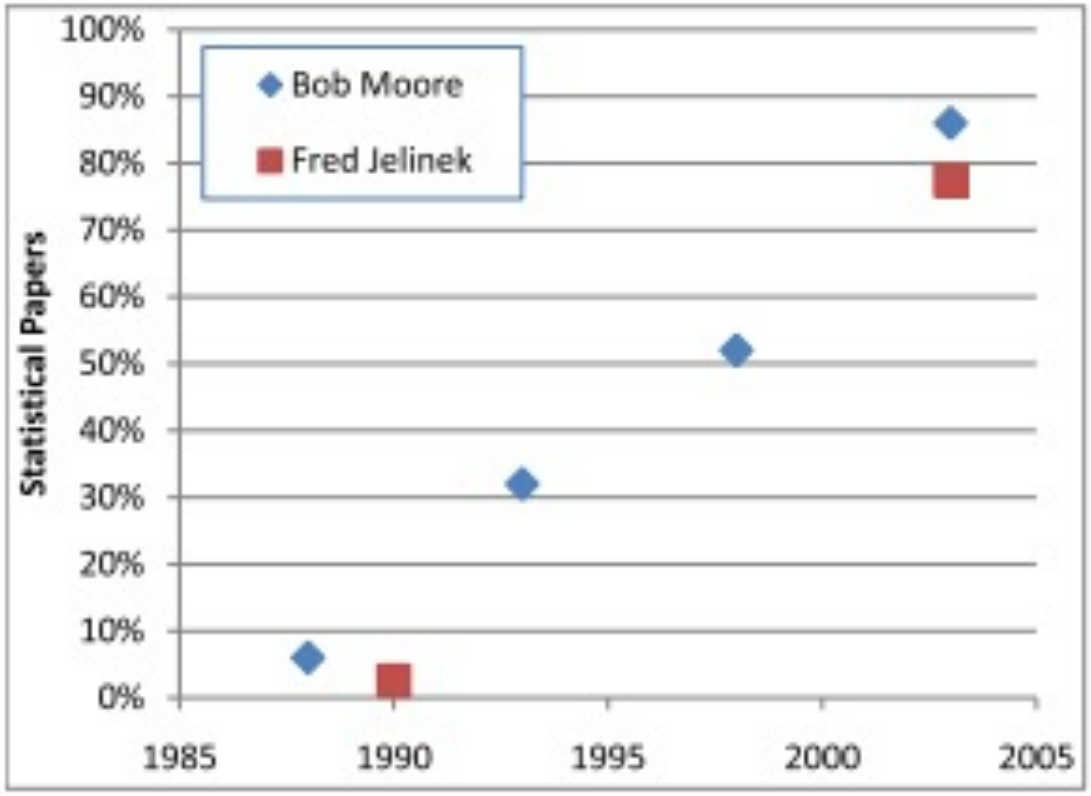


FIGURE 1 The shift from Rationalism to Empiricism is striking (and no longer controversial). This plot is based on two independent surveys of ACL meetings by Bob Moore and Fred Jelinek (personal communication).

Church K., LiLT Volume 2, Issue 4 May 2007

Representation documents: Bag of Words

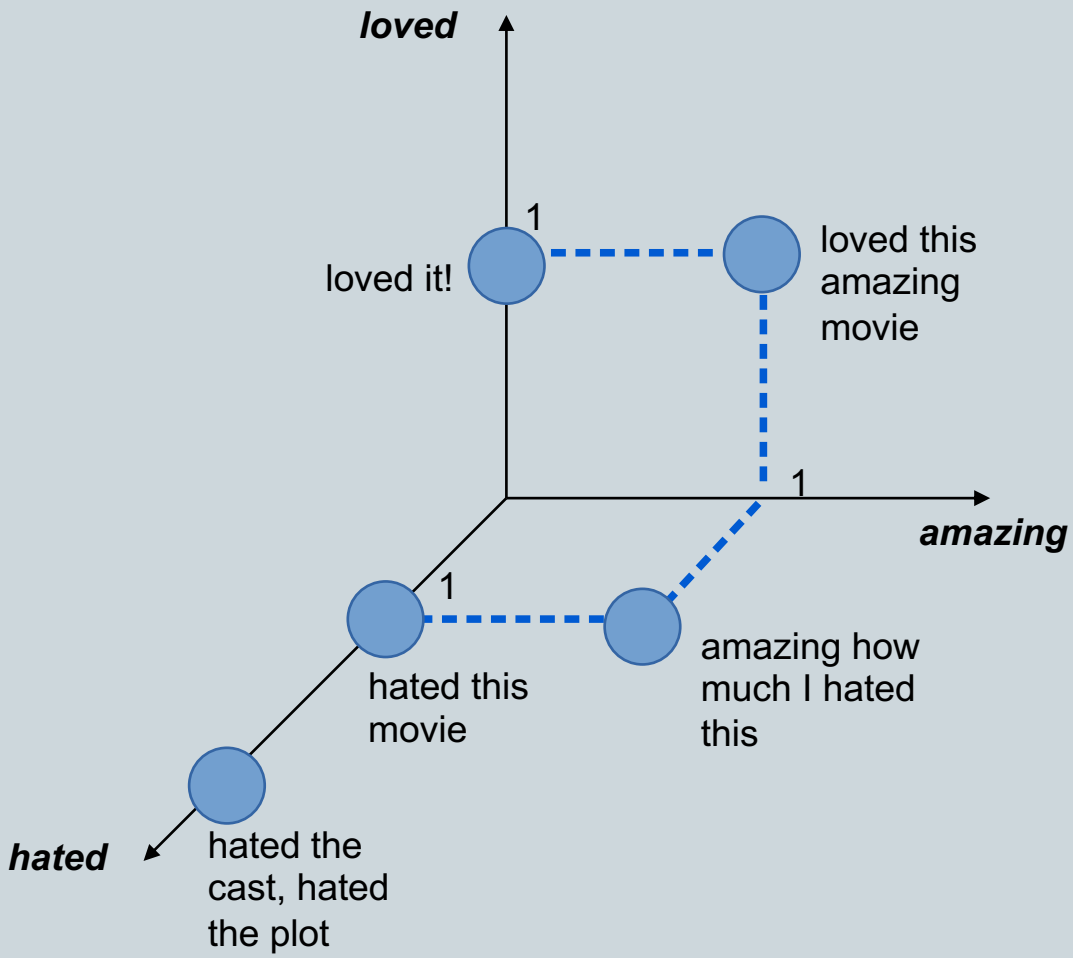
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

(Credit: Jurafsky and Martin, draft 3rd edition https://web.stanford.edu/~jurafsky/slp3/)

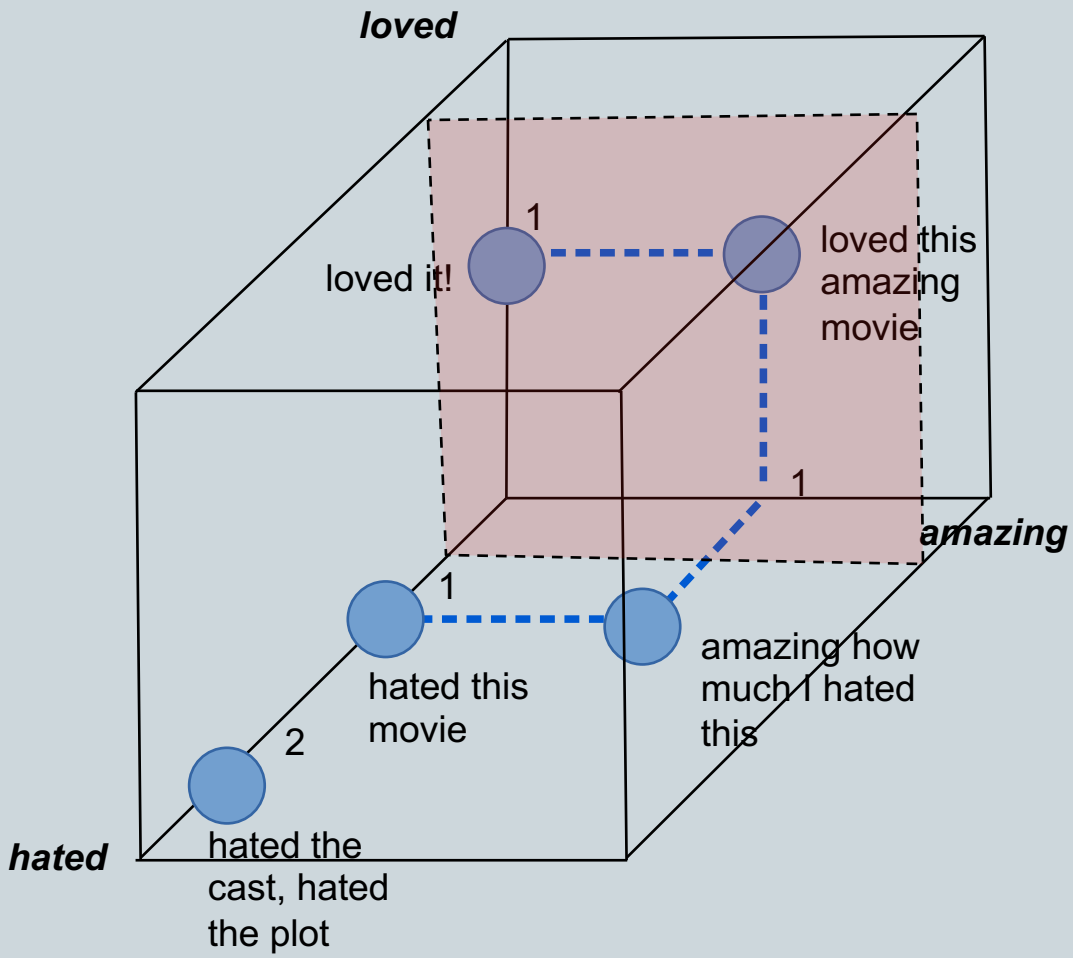
Bag of words as a vector space

- Here we have three dimensions in our model
- Typically, we would have hundreds or thousands of dimensions



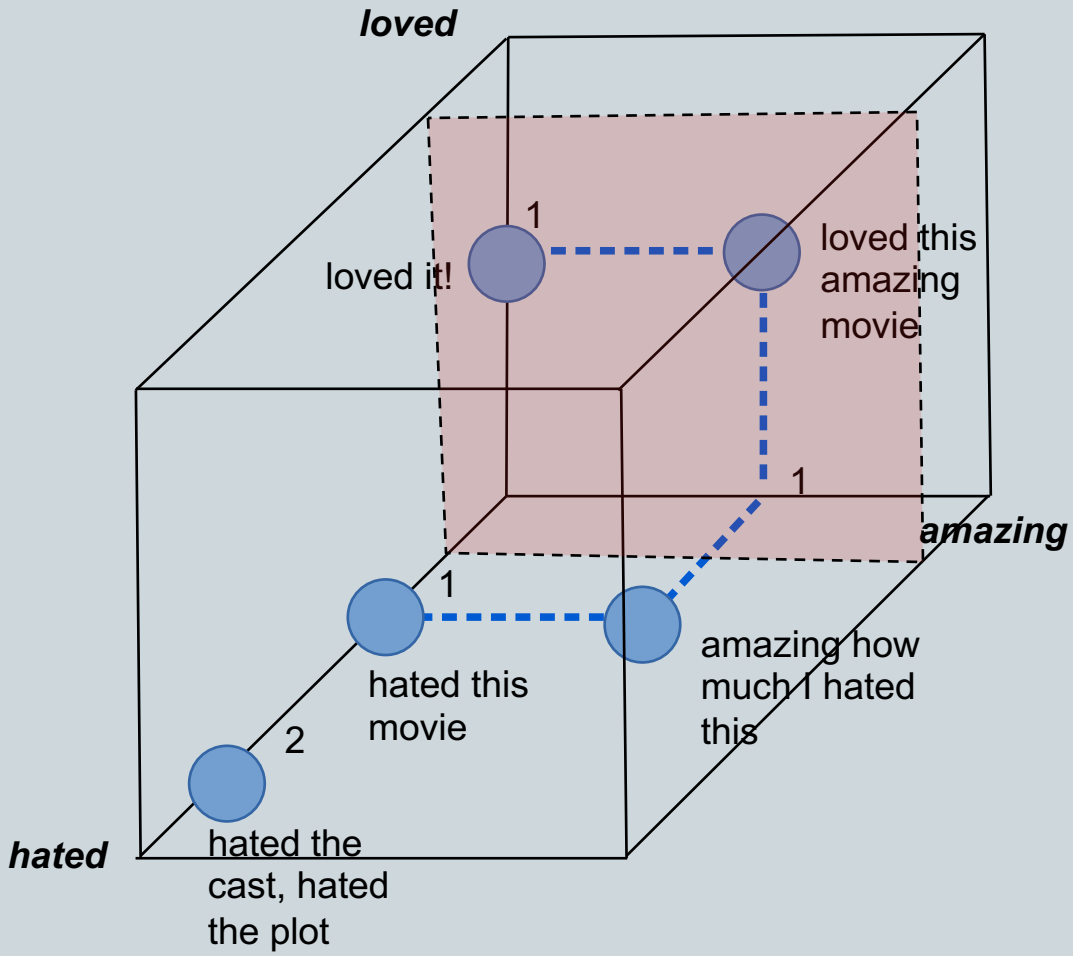
Bag of words

- We can create a plane in the ”feature space” to separate out different classes of our examples



Bag of words

- What about word order?
- What about negation?
- All words have the same influence, regardless of their salience
- More sophisticated representations are available!



Thank you

angus.roberts@kcl.ac.uk
https://www.kcl.ac.uk/people/angus-roberts