Topic Modeling

Matching Theory to Methods?

Topic Modeling: Example

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

We know this ‘corpus’ is structured from 2 topics, and we want to reverse engineer those two topics from the co-occurrence of words in each ‘document’.

Topic Modeling: Example

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

- Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
- Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)
- Sentences 1 and 2: 99% Topic A, 1% Topic B
- Sentences 3 and 4: 99% Topic B, 1% Topic A
- Sentence 5: 60% Topic A, 40% Topic B

Output: Topic Distribution over Words

Topic	broccoli	bananas	breakfast	kitten	cute	hamster	and	are	*Total*
A	.30	.25	.20	.01	.01	.01	.12	.10	1
B	.01	.01	.01	.35	.24	.25	.08	.05	1

Output: Document Distribution over Topics

Document	Topic A Weight	Topic B Weight	Total
1	.99	.01	1
2	.99	.01	1
3	.01	.99	1
4	.01	.99	1
5	.60	.40	1

Possible Output (takes an extra step)

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

Topic A (interpreted to be about Food)

Topic B (interpreted to be about Animals)

Caution!

Qualitative Decision Points:

- Which features? Weights?
  - Stop words? Capitalization? Punctuation?
- How many topics?
- Which algorithm?
- How do you adjudicate between all these choice points?

Caution!

Qualitative Decision Points:

- Which features? Weights?
  - Stop words? Capitalization? Punctuation?
- How many topics?
- Which algorithm?
- How do you adjudicate between all these choice points?

Topic models never devise the one, best way to cluster your corpus.

It is not objective, or perfect, or strictly scientific, etc.