3.10 Clustering Text

After we have calculated each center of clusters

For example, if we want to know which cluster doc1 belongs to:

- **Distance between doc1 with center1: 7**
- **Distance between doc1 with center2: 9**
- **Distance between doc1 with center3: 8**
- **Distance between doc1 with center4: 27**

Give doc1 the cluster1 with the smallest distance

Similarly,

- **Distance between doc2 with center1: 18**
- **Distance between doc2 with center2: 11**
- **Distance between doc2 with center3: 38**
- **Distance between doc2 with center4: 21**

	the	Word 2	…
My tweet	1	0	…
Declined and Fall	25,000	10	…

Present each word

Present each document

The correlations between words:

hitter baseball porcupine

Factor = [0.91, 0.82, …, 0], a = scores

Then we can use factor and a to represent each document

Doc1 = a_{11} factor 1 + a_{12} factor 2 + … + a_{1n}factor n
Doc1 = a_{21} factor 1 + a_{22} factor 2 + … + a_{2n}factor n

Dimension Reduction
(k << n)

Doc1 = a_{11} factor 1 + a_{12} factor 2 + … + a_{1k}factor k
Doc1 = a_{21} factor 1 + a_{22} factor 2 + … + a_{2k}factor k

$$TF = \ times\ t\ occurs$$

$$IDF = \frac{\#\ of\ docs}{\#\ where\ term\ t\ occurs}$$

$$\cos\theta = \frac{V1 \cdot V2}{\|V1\| \cdot \|V2\|} = \frac{\sum_{t=1}^{D} V1_t V2_t}{\left(\sqrt{\sum_{t=1}^{D} V1_t^2}\right)\left(\sqrt{\sum_{t=1}^{D} V2_t^2}\right)}$$

Doc 1 -> cluster 1
Doc 2 -> cluster 2
Doc 3 -> cluster 1
Doc 4 -> cluster 2

Doc 1,3 similar
Doc 2,4 similar

To describe cluster 1:
Mean = (a_1, …, a_{20})
Description = a_1(loading 1) + … + a_{20}(loading20)