Term weighting

tfidf

Term-document count matrices(计数矩阵)

- Consider：
  - the number of occurrences of a term in a document:
  - **Bag of words**(词袋) model
  - Document is a vector in $\mathbb{N}^{|V|}$: a column below

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Counts vs. frequencies(频率)

- Consider again the query : ***ides of march***
  - *Julius Caesar* has 5 occurrences of ***ides***
  - No other play has ***ides***
  - ***march*** occurs in over a dozen
  - All the plays contain ***of***
- By this scoring measure, the top-scoring play is likely to be the one with the most ***of***s
- "frequency" is used to mean "count" in IR

Term frequency *tf*

- Long docs are favored because they're more likely to contain query terms
- Can fix this to some extent by normalizing for document length
- But is **raw *tf*** the right measure?

Weighting term frequency: *tf*

- What is the relative importance of
  - 0 vs. 1 occurrence of a term in a doc
  - 1 vs. 2 occurrences
  - 2 vs. 3 occurrences …
- Unclear
  - while it seems that more is better, a lot isn't proportionally(比例) better than a few
  - Can just use raw ***tf***

$$wf_{t,d} = 0 \quad \text{if } tf_{t,d} = 0$$
$$wf_{t,d} = 1 + \log tf_{t,d} \quad \text{otherwise}$$

Score computation

- Score for a query ***q* = sum over terms *t* in *q***:

$$= \sum_{t \in q} tf_{t,d}$$

- [Note: 0 if no query terms in document]
- This score can be zone-combined
  - Can use *wf* instead of *tf* in the above
- **Still doesn't consider** : term scarcity(稀缺) in collection
  - ***ides*** is rarer than ***of***

Weighting should depend on the term overall

- Which of these tells you more about a doc?
  - 10 occurrences of *hernia*?
  - 10 occurrences of *the*?
- Would like to attenuate the weight of a common term
  - But what is "common"?
- Suggest looking at collection frequency (***cf***)
  - The total number of occurrences of the term in the entire collection of documents

Document frequency (*df*)

May be better:

- ***df* =** number of docs in the corpus containing the term

Word	*cf*	*df*
ferrari	10422	17
insurance	10440	3997

- Document/collection frequency weighting is only possible in known (static) collection.
- So how do we make use of ***df*** ?

tf x idf term weights

- **tf x idf** measure combines:
  - term frequency (***tf***) or weighting term frequency (***wf***)
    - some measure of term density in a doc
  - **inverse document frequency** (***idf***)
    - Measure of informativeness(信息含量) of a term
      - its rarity across the whole corpus
    - The most commonly used version is: $idf_i = \log\left(\frac{n}{df_i}\right)$
- **See** Kishore Papineni, NAACL 2, 2002 for theoretical justification

Summary: tf x idf or tf.idf

- Assign
  - a **tf.idf** weight to each term ***i*** in each document ***d***

$$w_{i,d} = tf_{i,d} \times \log(n / df_i)$$

$tf_{i,d}$ = frequency of term i in document j

n = total number of documents

df_i = the number of documents that contain term i

- Increases with the number of occurrences *within* a doc
- Increases with the rarity of the term *across* the whole corpus

Real-valued(真值) term-document matrices

- Function (scaling) of count of a **word** in a document:
  - **Bag of words** model
  - Each is a vector in $\mathbb{R}^v$
  - Here log-scaled ***tf.idf***

Note can be >1!

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	13.1	11.4	0.0	0.0	0.0	0.0
Brutus	3.0	8.3	0.0	1.0	0.0	0.0
Caesar	2.3	2.3	0.0	0.5	0.3	0.3
Calpurnia	0.0	11.2	0.0	0.0	0.0	0.0
Cleopatra	17.7	0.0	0.0	0.0	0.0	0.0
mercy	0.5	0.0	0.7	0.9	0.9	0.3
worser	1.2	0.0	0.6	0.6	0.6	0.0