

Querying

Introduction to Information Retrieval

INF 141/ CS 121

Donald J. Patterson

Content adapted from Hinrich Schütze

<http://www.informationretrieval.org>



Term Frequency Matrix

- Bag of words
- Document is vector with integer elements

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



Term Frequency - tf

- Long documents are favored because they are more likely to contain query terms
- Reduce the impact by normalizing by document length
- Is raw term frequency the right number?



Weighting Term Frequency - WTF

- What is the relative importance of
 - 0 vs. 1 occurrence of a word in a document?
 - 1 vs. 2 occurrences of a word in a document?
 - 2 vs. 100 occurrences of a word in a document?
- Answer is unclear:
 - More is better, but not proportionally
 - An alternative to raw tf:

```
WTF( $t, d$ )  
1  if  $tf_{t,d} = 0$   
2    then return(0)  
3    else return( $1 + \log(tf_{t,d})$ )
```



Weighting Term Frequency - WTF

$$WTF(t, d)$$

- The score for query, q , is
 - 1 **if** $tf_{t,d} = 0$
 - 2 **then** *return*(0)
 - 3 **else** *return*($1 + \log(tf_{t,d})$)
- Sum over terms, t

$$Score_{WTF}(q, d) = \sum_{t \in q} (WTF(t, d))$$

What is the score of “bill rights” in the declaration of independence?



Weighting Term Frequency - WTF

$$WTF(t, d)$$

- The score for query, q , is
 - Sum over terms, t
- $$\begin{array}{ll}
 1 & \text{if } tf_{t,d} = 0 \\
 2 & \text{then return}(0) \\
 3 & \text{else return}(1 + \log(tf_{t,d}))
 \end{array}$$

$$Score_{WTF}(q, d) = \sum_{t \in q} (WTF(t, d))$$

$$\begin{aligned}
 Score_{WTF}(\text{"bill rights"}, declarationOfIndependence) &= \\
 WTF(\text{"bill"}, declarationOfIndependence) &+ \\
 WTF(\text{"rights"}, declarationOfIndependence) &=
 \end{aligned}$$

$$0 + 1 + \log(3) = 1.48$$



Weighting Term Frequency - WTF

$$Score_{WTF}(q, d) = \sum_{t \in q} (WTF(t, d))$$

$$\begin{aligned} Score_{WTF}(\text{"bill rights"}, \text{declarationOfIndependence}) &= \\ WTF(\text{"bill"}, \text{declarationOfIndependence}) &+ \\ WTF(\text{"rights"}, \text{declarationOfIndependence}) &= \\ 0 + 1 + \log(3) &= 1.48 \end{aligned}$$

$$\begin{aligned} Score_{WTF}(\text{"bill rights"}, \text{constitution}) &= \\ WTF(\text{"bill"}, \text{constitution}) &+ \\ WTF(\text{"rights"}, \text{constitution}) &= \\ 1 + \log(10) + 1 + \log(1) &= 3 \end{aligned}$$



Weighting Term Frequency - WTF

- Can be zone combined:

$$\begin{aligned} \text{Score} = & 0.6(\text{Score}_{WTF}(\text{"instant oatmeal health"}, d.\text{title}) + \\ & 0.3(\text{Score}_{WTF}(\text{"instant oatmeal health"}, d.\text{body}) + \\ & 0.1(\text{Score}_{WTF}(\text{"instant oatmeal health"}, d.\text{abstract})) \end{aligned}$$

- Note that you get 0 if there are no query terms in the document.
- Is that really what you want?
- We will eventually address this



Unsatisfied with term weighting

- Which of these tells you more about a document?
 - 10 occurrences of “mole”
 - 10 occurrences of “man”
 - 10 occurrences of “the”
- It would be nice if common words had less impact
 - How do we decide what is common?
- Let's use **corpus-wide statistics**



Corpus-wide statistics

- **Collection Frequency**, cf
 - Define: The total number of occurrences of the term in the entire corpus
- **Document Frequency**, df
 - Define: The total number of documents which contain the term in the corpus



Corpus-wide statistics

<i>Word</i>	<i>Collection Frequency</i>	<i>Document Frequency</i>
-------------	-----------------------------	---------------------------

<i>insurance</i>	10440	3997
------------------	-------	------

<i>try</i>	10422	8760
------------	-------	------

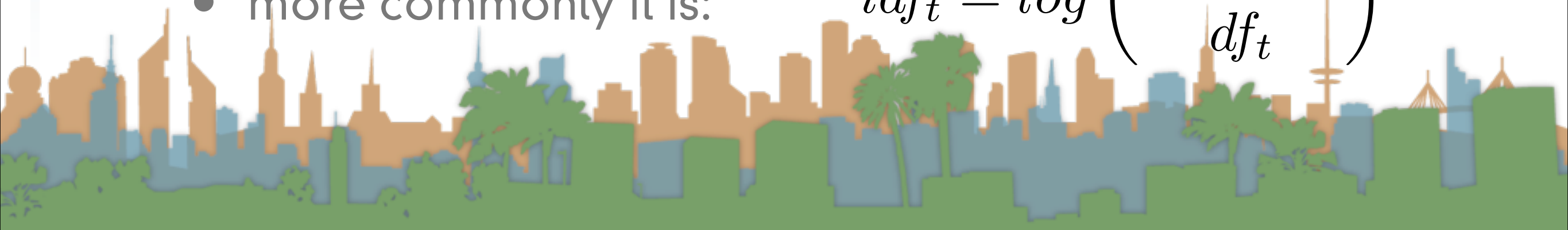
- This suggests that df is better at discriminating between documents
- How do we use df?



Corpus-wide statistics

- Term-Frequency, Inverse Document Frequency Weights
 - “tf-idf”
 - tf = term frequency
 - some measure of term density in a document
 - idf = inverse document frequency
 - a measure of the informativeness of a term
 - it's rarity across the corpus
 - could be just a count of documents with the term
 - more commonly it is:

$$idf_t = \log \left(\frac{|corpus|}{df_t} \right)$$

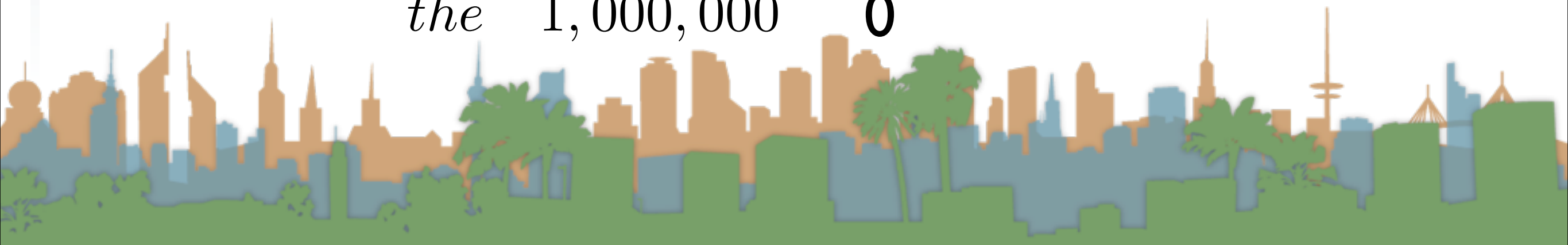


TF-IDF Examples

$$idf_t = \log \left(\frac{|corpus|}{df_t} \right)$$

$$idf_t = \log_{10} \left(\frac{1,000,000}{df_t} \right)$$

<i>term</i>	<i>df_t</i>	<i>idf_t</i>
<i>calpurnia</i>	1	6
<i>animal</i>	10	4
<i>sunday</i>	1000	3
<i>fly</i>	10,000	2
<i>under</i>	100,000	1
<i>the</i>	1,000,000	0



TF-IDF Summary

- Assign tf-idf weight for each term t in a document d :

$$tfidf(t, d) = WTF(t, d) * \log \left(\frac{|corpus|}{df_{t,d}} \right) (1 + \log(tf_{t,d}))$$

- Increases with number of occurrences of term in a doc.
- Increases with rarity of term across entire corpus
- Three different metrics
 - term frequency
 - document frequency
 - collection/corpus size



Now, real-valued term-document matrices

- Bag of words model
- Each element of matrix is tf-idf value

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

The numbers are just examples, they are not correct with respect to tf-idf and the previous slide

Vector Space Scoring

- That is a nice matrix, but
 - How does it relate to scoring?
 - Next, vector space scoring

