Vector Space Scoring

Introduction to Information Retrieval
INF 141
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Corpus-wide statistics

Corpus-wide statistics

- Collection Frequency, cf
  - Define: The total number of occurences of the term in the entire corpus

Corpus-wide statistics

- Collection Frequency, cf
  - Define: The total number of occurences 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

Word	*Collection Frequency*	*Document Frequency*
insurance	10440	3997
try	10422	8760

Corpus-wide statistics

Word	*Collection Frequency*	*Document Frequency*
insurance	10440	3997
try	10422	8760

- This suggests that df is better at discriminating between documents

Corpus-wide statistics

$Word$	$Collection\ Frequency$	$Document\ Frequency$
$insurance$	10440	3997
try	10422	8760

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

Corpus-wide statistics

Corpus-wide statistics

- Term-Frequency, Inverse Document Frequency Weights

Corpus-wide statistics

- Term-Frequency, Inverse Document Frequency Weights
  - “tf-idf”

Corpus-wide statistics

- Term-Frequency, Inverse Document Frequency Weights
  - “tf-idf”
  - tf = term frequency

Corpus-wide statistics

- Term-Frequency, Inverse Document Frequency Weights
  - "tf-idf"
  - tf = term frequency
    - some measure of term density in a document

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

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

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

Corpus-wide statistics

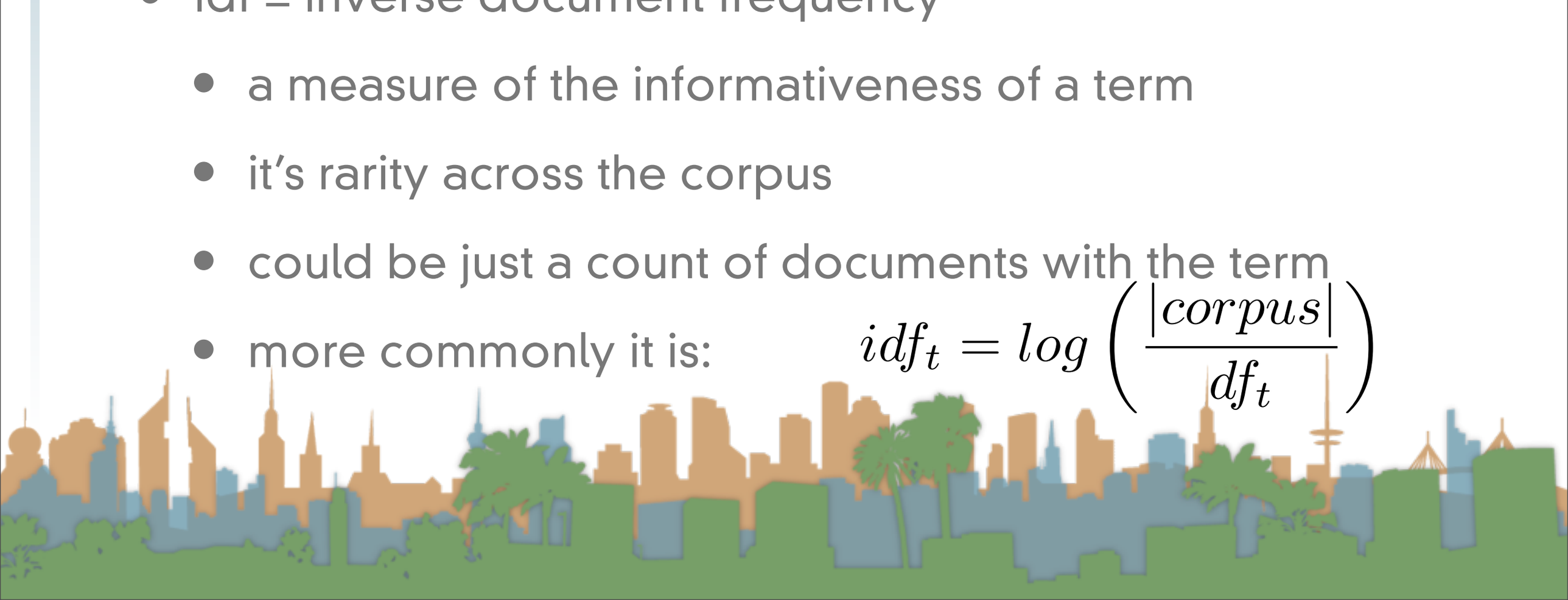
- Term-Frequency, Inverse Document Frequency Weights
  - “tf-idf”
  - tf = term frequency
    - some measure of term density in a document
  - idf = inverse document frequency
    - a measure of the informativeness of a term
    - it’s rarity across the corpus
    - could be just a count of documents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

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:

Corpus-wide statistics

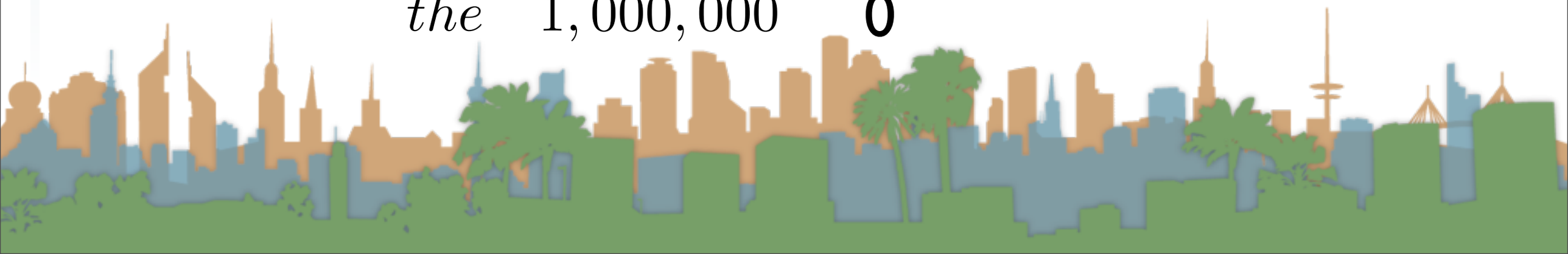
- Term-Frequency, Inverse Document Frequency Weights
  - "tf-idf"
  - tf = term frequency
    - some measure of term density in a document
  - idf = inverse document frequency
    - a measure of the informativeness of a term
    - it's rarity across the corpus
    - could be just a count of documents with the term
    - more commonly it is: $idf_t = log\left(\frac{|corpus|}{df_t}\right)$



TF-IDF Examples

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

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



TF-IDF Summary

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

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

  - 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 frequency

Now, real-valued term-document matrices

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

	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

Vector Space Scoring

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

Vector Space Model

- Define: Vector Space Model
  - Representing a set of documents as vectors in a common vector space.
  - It is fundamental to many operations
    - (query,document) pair scoring
    - document classification
    - document clustering
- Queries are represented as a document
  - A short one, but mathematically equivalent

Vector Space Model

- Define: Vector Space Model
  - A document, d, is defined as a vector: $\vec{V}(d)$
    - One component for each term in the dictionary
    - Assume the term is the tf-idf score

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

- A corpus is many vectors together.
- A document can be thought of as a point in a multi-dimensional space, with axes related to terms.

Vector Space Model

- Recall our Shakespeare Example:

	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

Vector Space Model

- Recall our Shakespeare Example:

$\vec{V}(d_1)$

	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

Vector Space Model

- Recall our Shakespeare Example:

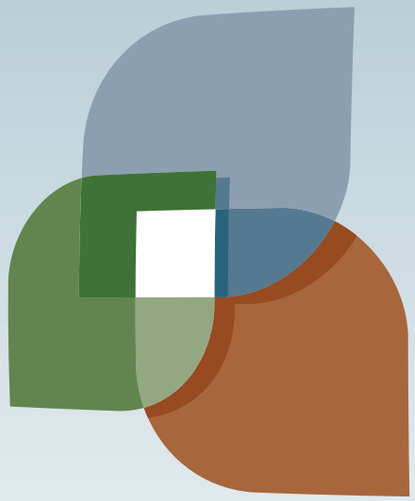
	$\vec{V}(d_1)$	$\vec{V}(d_2)$				$\vec{V}(d_6)$
	Antony and Cleopatra	*Julius Caesar*	*The Tempest*	*Hamlet*	*Othello*	*Macbeth*
Antony	13.1	11.4	0.0	0.0	0.0	0.0
Brutus	3.0	8.3	0.0	1.0	0.0	0.0
Caesar	2.3	2.3	0.0	0.5	0.3	0.3
Calpurnia	0.0	11.2	0.0	0.0	0.0	0.0
Cleopatra	17.7	0.0	0.0	0.0	0.0	0.0
mercy	0.5	0.0	0.7	0.9	0.9	0.3
worser	1.2	0.0	0.6	0.6	0.6	0.0

Vector Space Model

- Recall our Shakespeare Example:

	$\vec{V}(d_1)$	$\vec{V}(d_2)$				$\vec{V}(d_6)$
	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

$\vec{V}(d_6)_7$



LUCI