

# Day 4: Text Analysis

# Relevant Terms

using TF-IDF

# Intuition

- Count # times each word is used

TF

We are in the process of trying to arrange a conference call with you on either Tuesday or Wednesday of next week to discuss the paper which is attached

We will be doing this by conference call and once we set a time to talk with you, will give you the number to call.

# Intuition

- Count # times each word is used

TF

We are in the process of trying to arrange a **conference** call with you on either Tuesday or Wednesday of next week to discuss the paper which is attached.

We will be doing this by **conference** call and once we set a time to talk with you, will give you the number to call.

**conference: 2**

# Intuition

- Count # times each word is used

TF

We are in the process of trying **to** arrange a conference call with you on either Tuesday or Wednesday of next week **to** discuss the paper which is attached.

We will be doing this by conference call and once we set a time **to** talk with you, will give you the number **to** call.

**to: 4**

# Intuition

- Count # times each word is used **TF**
- Penalize if most documents use word **IDF**

We are in the process of trying **to** arrange a conference call with you on either Tuesday or Wednesday of next week **to** discuss the paper which is attached.

We will be doing this by conference call and once we set a time **to** talk with you, will give you the number **to** call.

**to: 4** ← Want to penalize

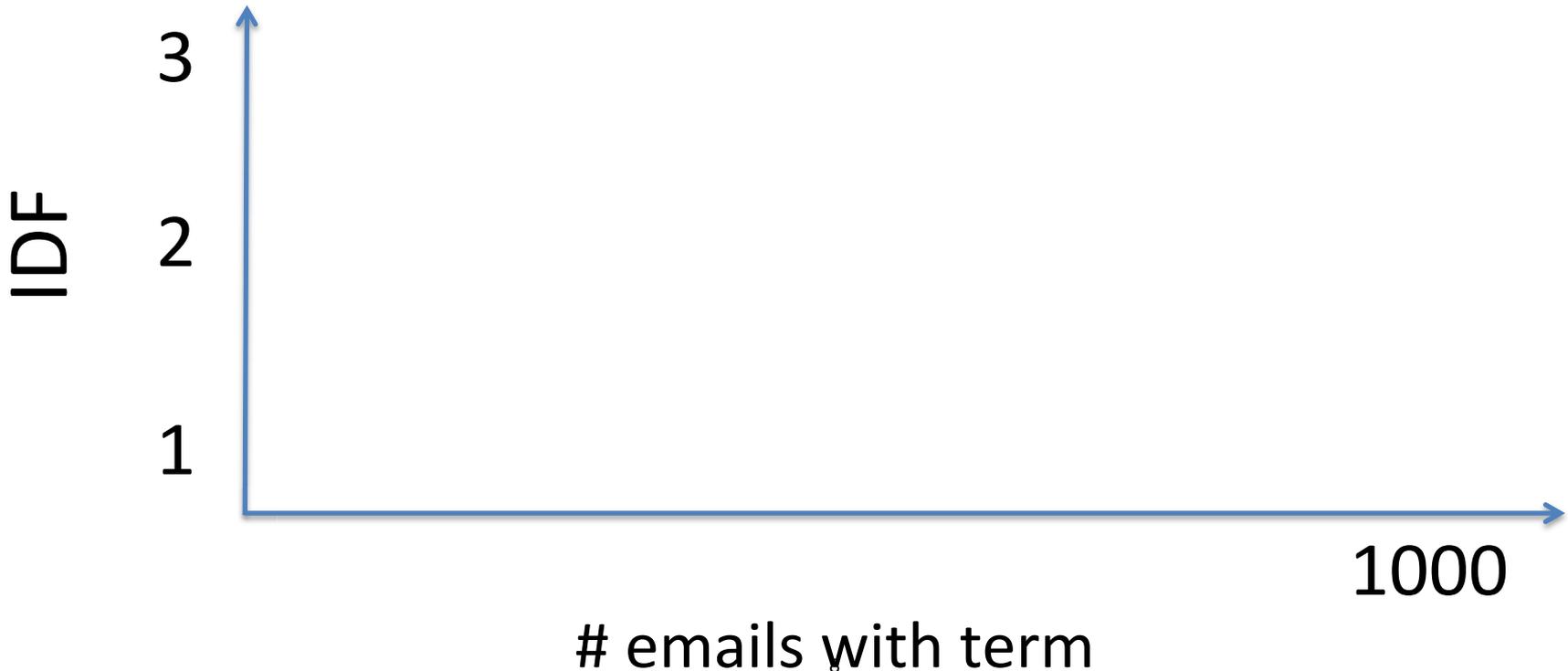
# IDF. How do they work?

$\log(\text{total \# emails} / \text{\# emails with word})$

# IDF. How do they work?

$$\log(\text{total \# emails} / \text{\# emails with word})$$

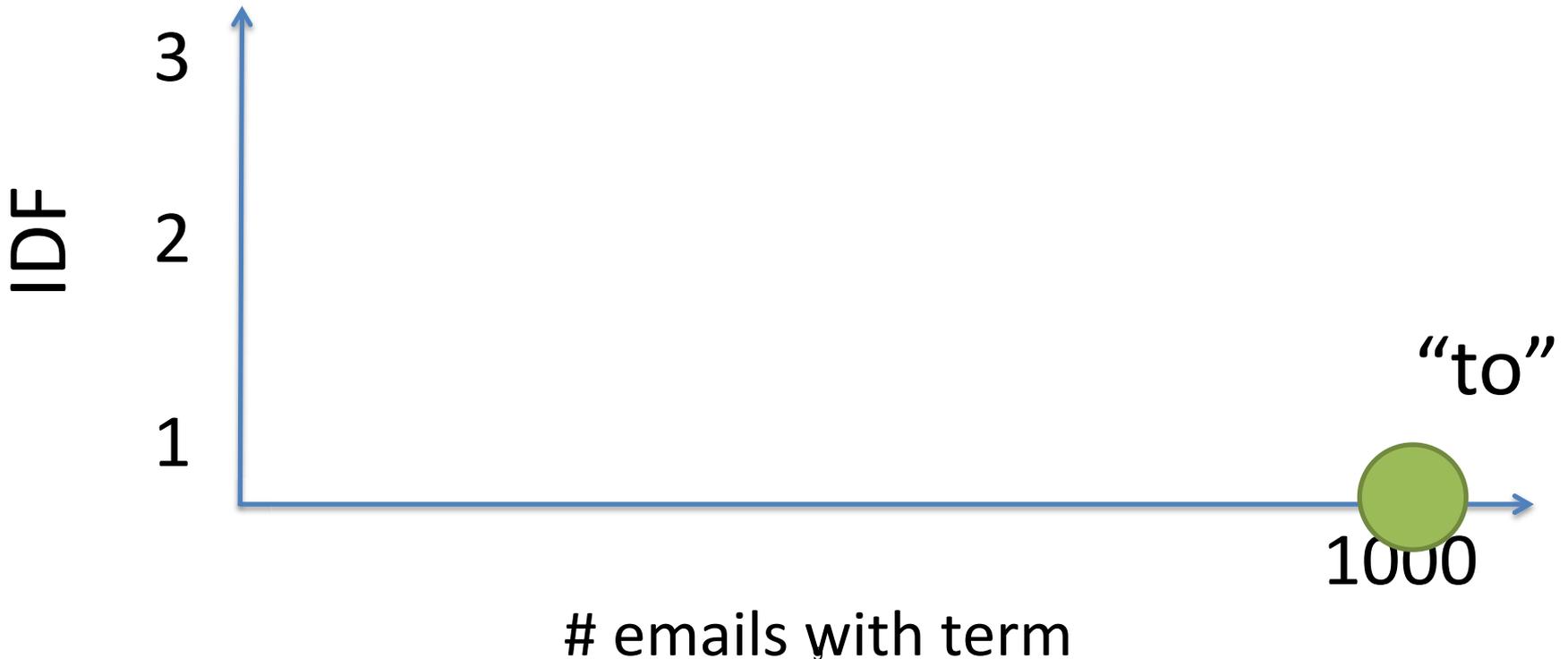
1000 Total Emails



# IDF. How do they work?

$$\log(\text{total \# emails} / \text{\# emails with word})$$

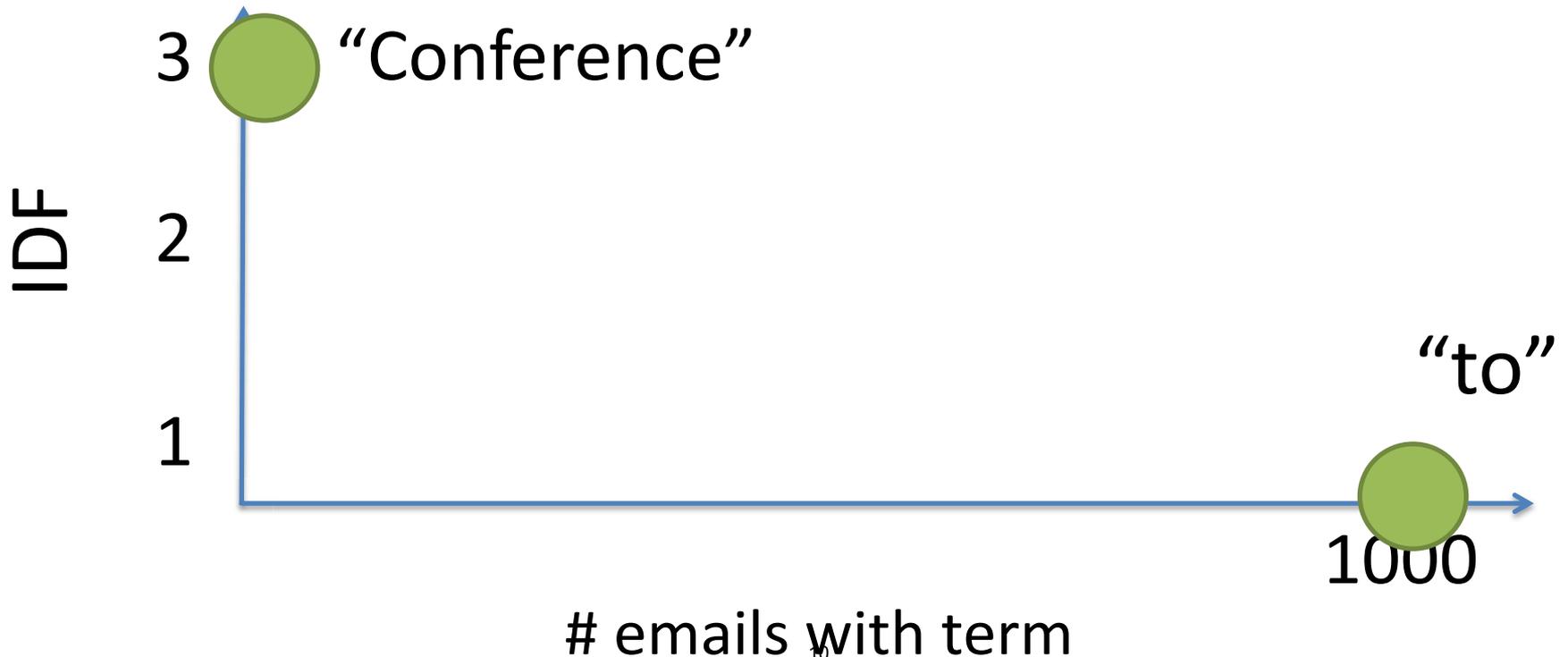
1000 Total Emails



# IDF. How do they work?

$$\log(\text{total \# emails} / \text{\# emails with word})$$

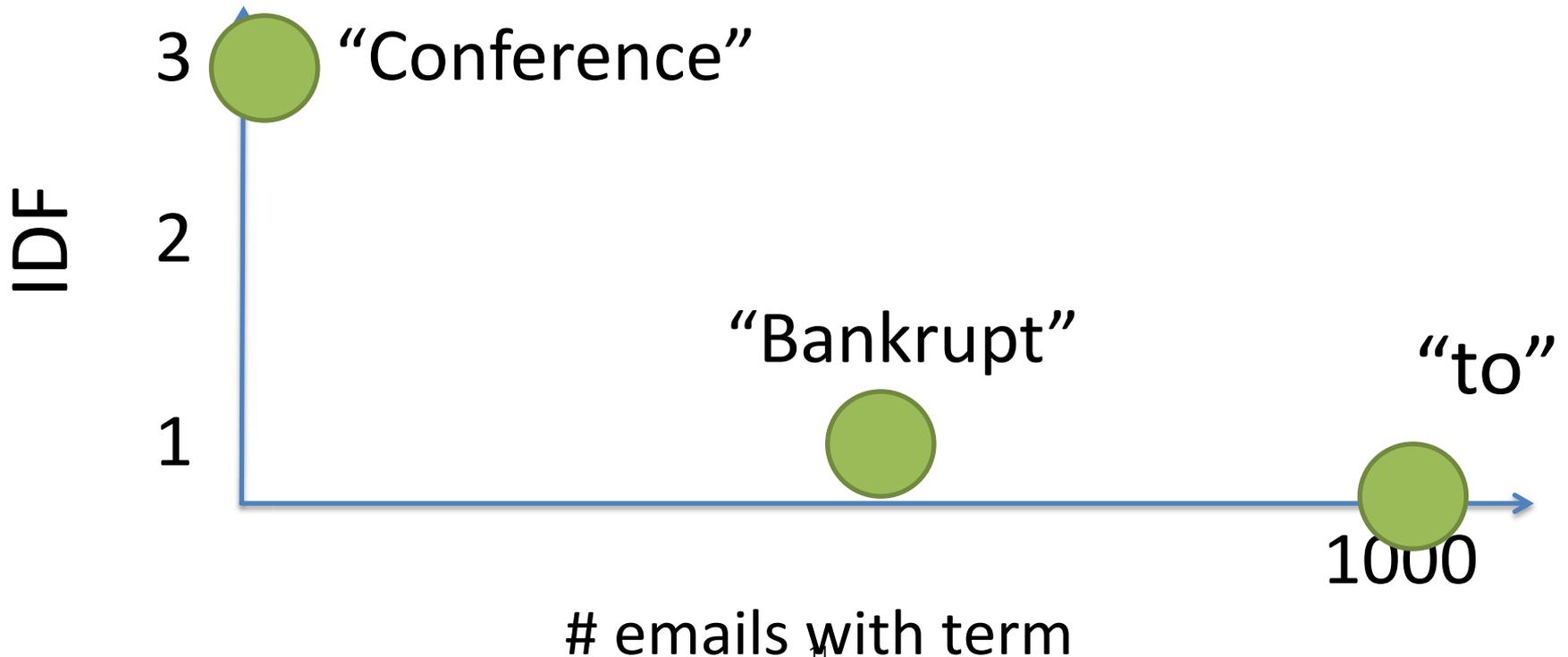
1000 Total Emails



# IDF. How do they work?

$$\log(\text{total \# emails} / \text{\# emails with word})$$

1000 Total Emails



# Relevant Terms

Frequent  
Words in email

But not in all  
emails

# Relevant Terms

Frequent  
Words in email

But not in all  
emails

TF

\*

IDF

# Relevant Terms

Frequent  
Words in email

But not in all  
emails

TF  
↑  
Per-Email

\*

IDF

↑  
Corpus Wide

# Relevant Terms

We are in the process of trying to arrange a **conference** call with you on either Tuesday or Wednesday of next week to discuss the paper which is attached.

We will be doing this by **conference** call and once we set a time to talk with you, will give you the number to call.

# Relevant Terms

We are in the process of trying to arrange a **conference** call with you on either Tuesday or Wednesday of next week to discuss the paper which is attached.

We will be doing this by **conference** call and once we set a time to talk with you, will give you the number to call.



	TFIDF
conference	4.1
call	3
...	
to	0.03
a	0.0012

How similar is email 1 to email  
2?

Cosine Similarity

Email1 = conference, enron, donuts,...

Email2 = enron, call, appointment,...

Email1 = conference, **enron**, donuts,...

Email2 = **enron**, call, appointment,...

Email1  $\cap$  Email2

Email1 = conference, **enron**, donuts,...

Email2 = **enron**, call, appointment,...

Email1  $\cap$  Email2

---

# words in both emails

$$\frac{\text{Email1} \bullet \text{Email2}}{\|\text{Email1}\| * \|\text{Email2}\|}$$

E1 = { 'conference' : 4, 'enron' : 3 }  
E2 = { 'enron' : 1, 'call' : 2 }

E1 = { 'conference' : 4, 'enron' : 3 }

E2 = { 'enron' : 1, 'call' : 2 }

E1[ 'enron' ] \* E2[ 'enron' ] + ...

E1 = { 'conference' : 4, 'enron' : 3 }

E2 = { 'enron' : 1, 'call' : 2 }

$$\frac{E1[ 'enron' ] * E2[ 'enron' ] + \dots}{\underbrace{\text{sqrt}( 4^2+3^2 )}_{E1} * \underbrace{\text{sqrt}( 1^2+2^2 )}_{E2}}$$

MIT OpenCourseWare  
<http://ocw.mit.edu>

Resource: How to Process, Analyze and Visualize Data  
Adam Marcus and Eugene Wu

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.