Day 4: Text Analysis


ENRON

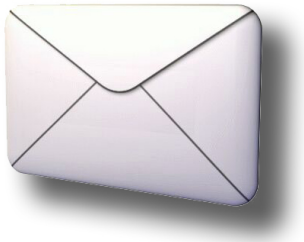
Kenneth Lay

- lay-k/
  - _sent/
  - all_documents/
  - business/
  - deleted_items/
  - family/
  - inbox/
  - sec_panel/
  - sent/

- lay-k/
  - _sent/
  - all_documents/
  - business/ → Email1, Email2, ….
  - deleted_items/
  - family/
  - inbox/
  - sec_panel/
  - sent/

2

- Relevant terms
  - Summaries
  - Searching

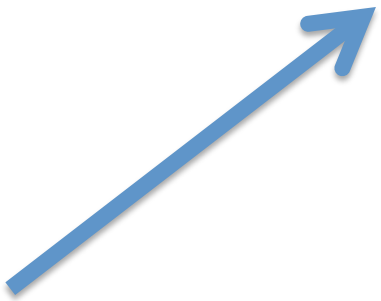


“Conference”
“Call”
“Saturday”
“Noon”

- Relevant terms
  - Summaries
  - Searching
- Similarity

- Relevant terms
  - Summaries
  - Searching
- Similarity

- Relevant terms
  - Summaries
  - Searching
- Similarity



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

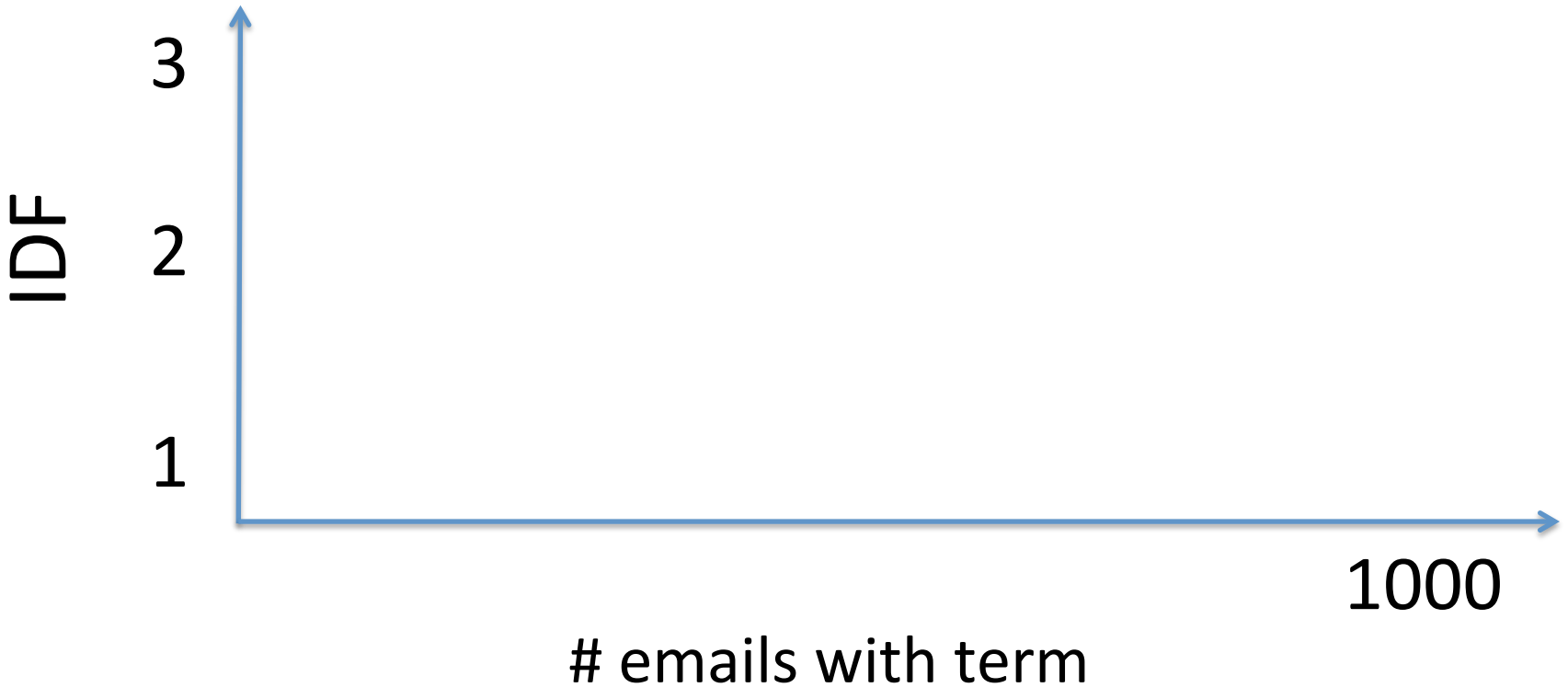
IDF. How do they work?

log(total # emails / # emails with word)

IDF. How do they work?

log(total # emails / # emails with word)

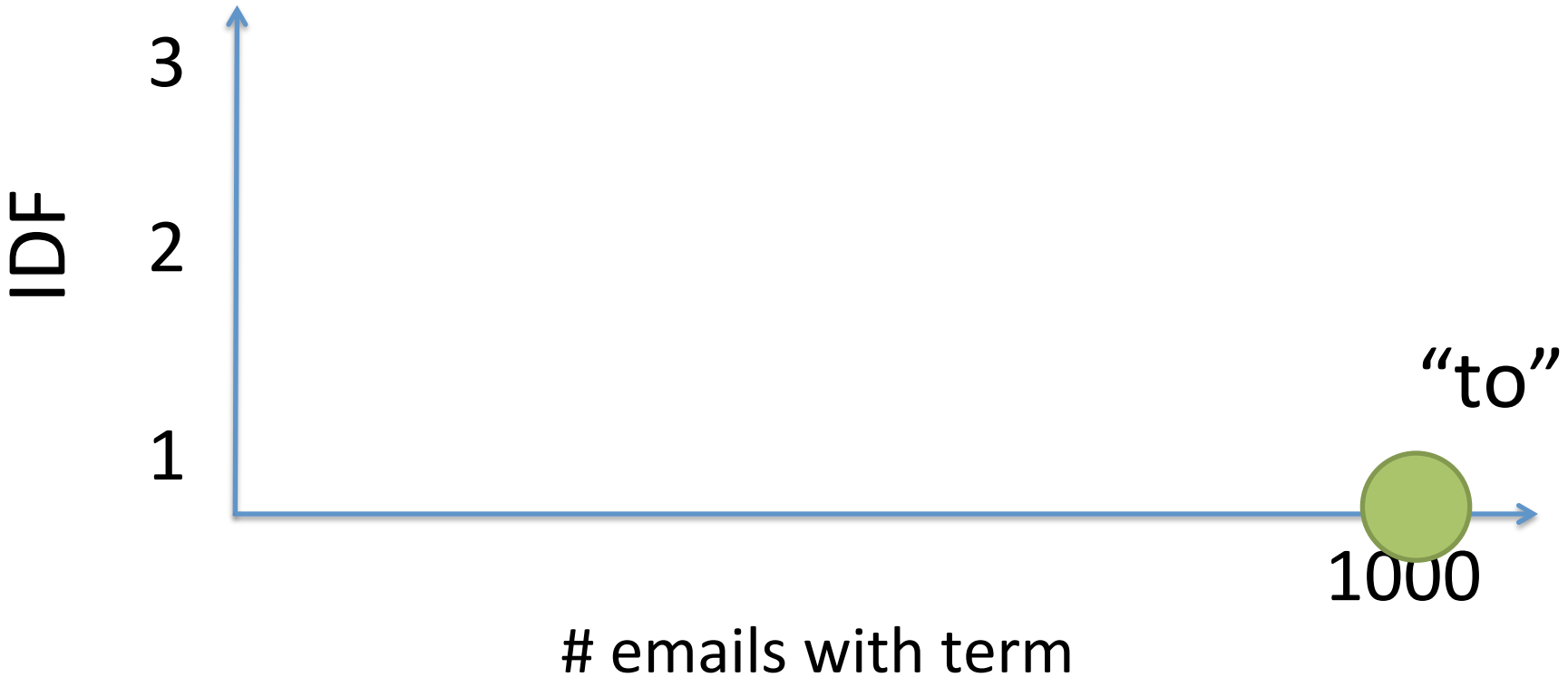
1000 Total Emails



IDF. How do they work?

log(total # emails / # emails with word)

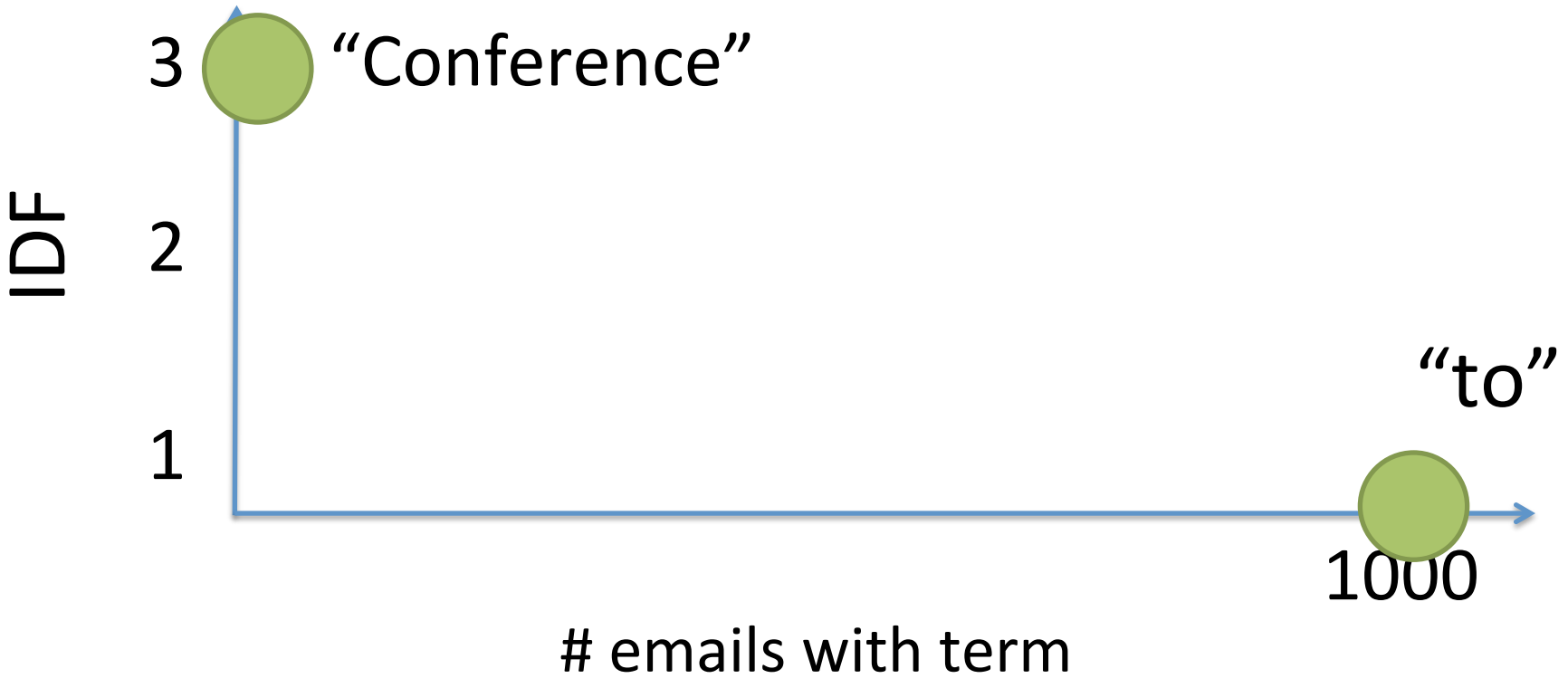
1000 Total Emails



IDF. How do they work?

log(total # emails / # emails with word)

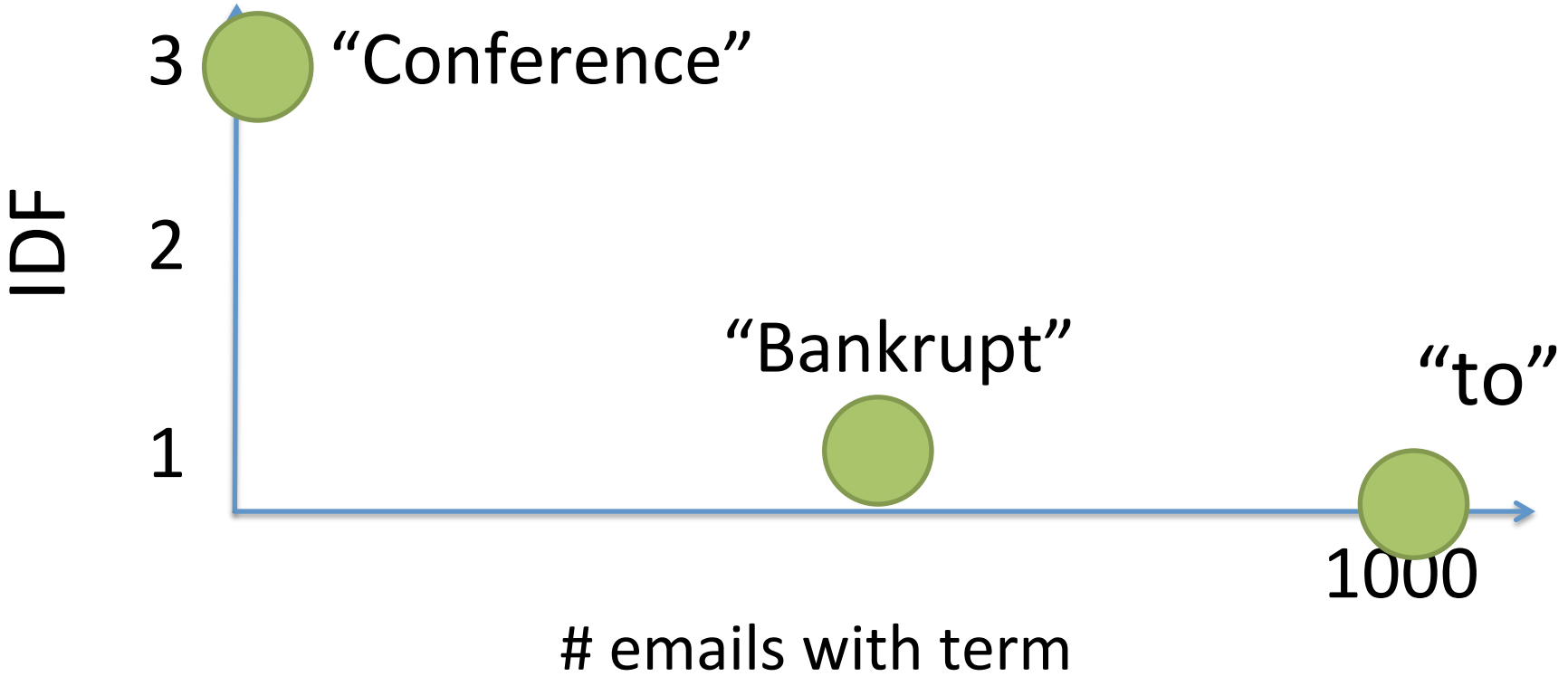
1000 Total Emails



IDF. How do they work?

log(total # emails / # 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

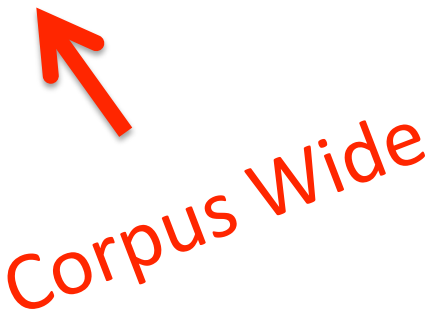
TF * IDF

Relevant Terms

Frequent Words in email

But not in all emails

TF * IDF



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,…

$$\text{Email1} \cap \text{Email2}$$

Email1 = conference, **enron**, donuts,…

Email2 = **enron**, call, appointment,…

$$\frac{\text{Email1} \cap \text{Email2}}{\#\ \text{words in both emails}}$$

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

```
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}
```

$$\frac{\texttt{E1[‘enron’]} * \texttt{E2[‘enron’]} + \ldots}{\texttt{sqrt}(\ 4^2+3^2\ ) * \texttt{sqrt}(\ 1^2+2^2\ )}$$

E1 E2

Data Cleaning

ok. 10:40 to be safe:P

On Fri, Jan 6, 2012 at 5:15 PM, Eugene Wu <sirrice@gmail.com> wrote:

> i have a feeling that breakfast foods stop at 11 at clover because thats
> how it works at the truck.
> so maybe 1045 or something is better.
>
>
> On Fri, Jan 6, 2012 at 5:09 PM, Adam Marcus <marcua@csail.mit.edu> wrote:
>
>> see you there at 11!
>>
>>
>> On Fri, Jan 6, 2012 at 5:05 PM, Eugene Wu <sirrice@gmail.com> wrote:
>>
>>> lets have brunch at 11. That way we skip the rush as well.
>>>

<html><body><img src="http://Link.p0.com/1x1.dyn?0VkG5BnoiRicmh9RA0Vs1eiXF=0" width=1 height=1 width="1" height="1" border="0"><table width="100%" cellpadding="0" cellspacing="0" border="0"><tr><td bgcolor="#f3f7fe"><table cellpadding="10" bgcolor="f3f7fe" border="0" cellspacing="0"><tr><td align="left"> View mobile version <table border="0" cellspacing="0" cellpadding="0" width="580"><tr><td width="1" height="

- Regular Expressions — “Enron’s” ~~“sirrice@gmail.com”~~
- Normalization — “Word” == “word”
- Stop Words — Ignore “a”, “to”, …

So we implemented all of that…

Sucked

Email Body vs Subject

Garbage in garbage out

- Forwarded message
- Unsubscribe
- Love
- Sincerely

Subject Line

- Fwd: 11/5 Dinner @ Jewel Bako. 7:45PM
- Dinner
- What time for dinner tomorrow?

Pretty Long

(optional)

Download your IMAP Email!

dataiap/resources/download_emails.py

Presentations Wednesday!

- 1 minute
- 2 slides by Tuesday night
- Google presentations

Interesting dataset

Interesting analysis

Topic Extensions

http://dataiap.github.com/dataiap/day4

git pull

- Remind them about
  - Downloading their own email
  - Last day presentations.
- Proton pump inhibitor