Lab for Media Search

National University of Singapore

Social Media Computing

Lecture 2: Text Processing

Lecturer: Aleksandr Farseev

E-mail: farseev@u.nus.edu

Slides: http://farseev.com/ainlfruct.html

Contents

- **What is Microblog**
- **Text Preprocessing**
- **Textual Data Representation**
- **Summary**

Blogging & Microblogging?

What is a blog?

- A **blog** (a portmanteau of the term "**web log**") is a type of website or part of a website.
  - Blogs are usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video.
  - Entries are commonly displayed in reverse-chronological order.

- Blog Resources

1. Go to http://en.wikipedia.org/wiki/Glossary_of_blogging.
   - Search for a definition of **video**, **audio** and **photo** blogs.
2. Use Blog Search Engine to find interesting Blogs (http://www.blogsearchengine.org/)
   - Find interesting blogs on the topic of Singapore?

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article

Interaction
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Donate to Wikipedia
Help

Toolbox

Print/export

Languages
తెలుగు

Glossary of blogging

From Wikipedia, the free encyclopedia

This article **may contain original research**. Please improve it by verifying the claims made and adding references. Statements consisting only of original research may be removed. More details may be available on the talk page. *(September 2007)*

This is a **list of blogging terms**. Blogging, like any hobby, has developed something of a specialised vocabulary. The following is an attempt to explain a few of the more common phrases and words, including etymologies when not obvious.

Contents

Blog-related terms [edit]

A [edit]

Atom
Another popular feed format developed as an alternative to RSS.

Autocasting
Automated form of podcasting that allows bloggers and blog readers to generate audio versions of text blogs from RSS feeds.

Audioblog
A blog where the posts consist mainly of voice recordings sent by mobile phone, sometimes with some short text message added for metadata purposes. (cf. podcasting)

B [edit]

Blawg
A law blog.

Bleg
An entry in a blog requesting information or contributions.

Blog Carnival
A blog article that contains links to other articles covering a specific topic. Most blog carnivals are hosted by a rotating list of frequent contributors to the carnival, and serve to both generate new posts by contributors and highlight new bloggers posting matter in that subject area.

Blog client
(weblog client) is software to manage (post, edit) blogs from operating system with no need to launch a web browser. A typical blog client has an editor, a spell-checker and a few more options that simplify content creation and editing.

Blogger
Person who runs a blog. Also blogger.com, a popular blog hosting web site. Rarely: weblogger.

Bloggernacle
Blogs written by and for Mormons (a portmanteau of "blog" and "Tabernacle)". Generally refers to faithful Mormon bloggers and sometimes refers to a specific grouping of faithful Mormon bloggers.

BlogSearchEngine.org

Search!

About 89,900,000 results (0.18 seconds)

Ads by Google related to: **singapore**

Agoda **Singapore** Hotels
www.agoda.com/**Singapore**_Hotels
Best Deals Today, Up to 80% Off Don't Miss Out, Book Now and Save!
971,820 people follow agoda on Google+

Singapore Hotels | 3 Star Hotels
Book Now | 5 Star Hotels
4 Star Hotels | Budget Hotels

★★★★★ rating for agoda.com

Managed Mobility Tax Free
www.crosspoint-telecom.com/
Upgrade your IP PBX Infrastructure Enjoy PIC Tax deductions today!

Singapore Blog Awards 2014 / 新加坡部落格大奖2014
Singapore Blog Awards returns for the 7th year with over $30000 worth of prizes to be won! Giddy-up now!
sgblogawards.omy.sg/

Expats In **Singapore** Blogs Directory
Expats In **Singapore** Blogs Directory at Expats Blog. Find a blog written by expatriates in **Singapore**, living and working in **Singapore**.
www.expatsblog.com/blogs/**singapore**

Singapore Dissident
1 day ago ... Ladies and Gentlemen, The brave **Singaporean** young man Roy Ngerng is being sued by Lee Kuan Yew's son for criticism on how the state ...
singaporedissident.blogspot.com/

Miss Tam Chiak: **Singapore** food blog
13 hours ago ... Miss Tam Chiak is a food and travel blog that features local food places in **Singapore** as well as around the world.
www.misstamchiak.com/

ieatishootipost: **Singapore** Food Reviews and Recipes
Get **Singapore's** best hawker and restaurant food reviews, recommendations and local food recipes. Never waste calories on yucky food!
ieatishootipost.sg/

PlayStation Blog - Asia | **Singapore** : The Official PlayStation Blog for ...
Hello everyone. Today, I am excited to announce two new titles are coming to Asia, The Last of Us™ Game of the Year Edition for PlayStation®3 and ...
blog.asia.playstation.com/community/sg/en

Singapore Law Blog

Examples of blog tasks

(adapted from Murray and Hourigan 2008)

Single-authored blogs

- Author's individual voice
- Creativity
- Reflective
- Vanity publishing factor
- Potential collaboration between student and teacher

Group blogs

- Collective dissemination of knowledge
- Peer discussion
- Collaborative processing and application of data
- Single publication: plurality of authors

Options to Create your own Blogs

- The best, easiest and most popular (free) options:

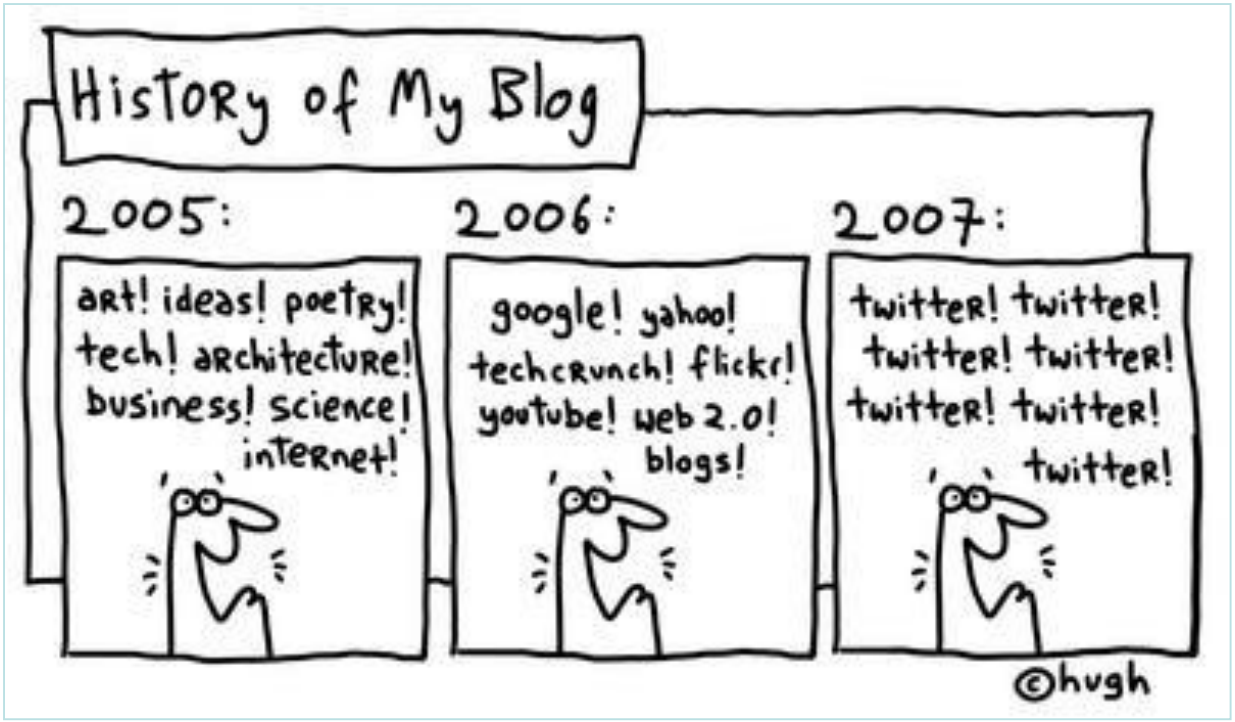
– www.blogger.com

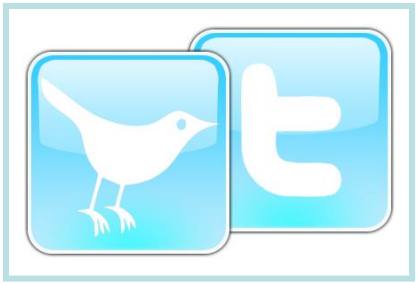
– www.edublogs.org

– www.wordpress.com

- Take your time to explore the interfaces and functionalities of these systems…

Influence of microblogging





What is microblogging?

- **Microblogging** is a form of blogging.
- A microblog differs from a traditional blog in that its content is typically much smaller, in both actual size and aggregate file size.
- A microblog entry could consist of nothing but a short sentence fragment, or an image or embedded video.

- See this Youtube video about microblogging (twitter):
  http://www.youtube.com/watch?v=ddO9idmax0o

Some microblogging sites

- Twitter (most popular)
- Edmodo (educationally oriented)
- Tumblr
- Jaiku
- ShoutEm
- among many others…

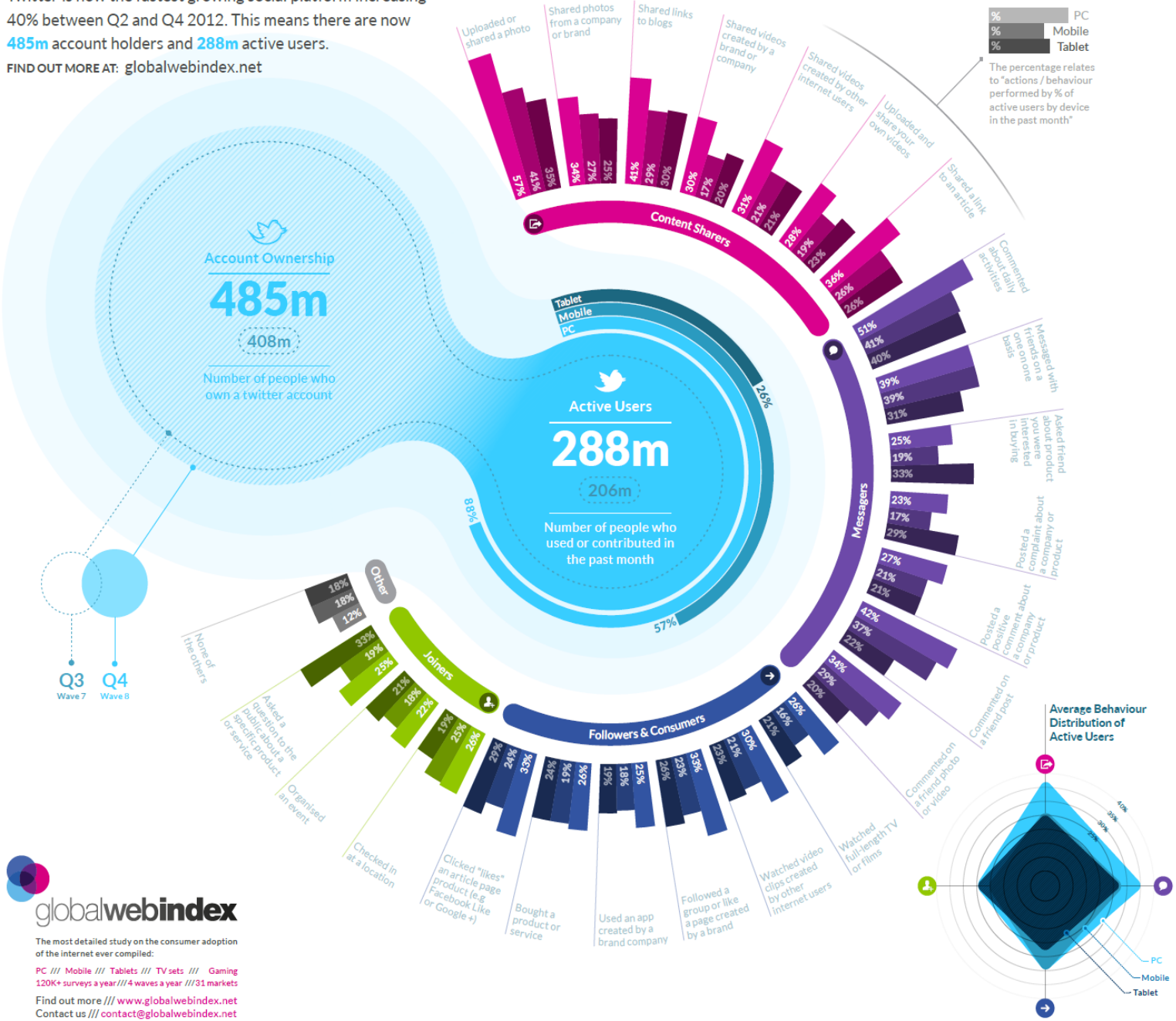
What's in a microblog?

Easy to share status messages

TWITTER The Fastest Growing Social Platform

Twitter is now the fastest growing social platform increasing 40% between Q2 and Q4 2012. This means there are now **485m** account holders and **288m** active users.

FIND OUT MORE AT: globalwebindex.net



globalwebindex

The most detailed study on the consumer adoption of the internet ever compiled:

PC /// Mobile /// Tablets /// TV sets /// Gaming
120K+ surveys a year /// 4 waves a year /// 31 markets

Find out more /// www.globalwebindex.net
Contact us /// contact@globalwebindex.net

Why so popular?

- Combines aspects of ***social networking*** **with aspects of** ***blogging***.

- Ambient Intimacy:

  "**Ambient intimacy** is about being able to ***keep in touch*** with people with a level of ***regularity and intimacy*** that you wouldn't usually have access to, because time and space conspire to make it impossible. "

  - Leisa Reichelt

.

What do people use Twitter for?

- Using Link Structure:
  - Information source

    Have a large number of followers (include bots like forecast, stock, CNN breaking news, etc.)
  - Information seeker

    Post infrequently, but have a number of connections
  - Friendship relation

    Most user's social network is within mutual acquaintances
- Using Content:
  - Daily chatter dinner, work, movie…
  - Conversations (@) Reply to a specific person @evgeniy
  - Sharing URLs Sharing URLs through tinyURL etc.
  - Commenting on News Number of automated RSS to Twitter bots posting news

Contents

- What is Microblog
- Text Preprocessing
- Textual Data Representation
- Summary

Tweets vs. Documents

From content aspect:

- Short vs. Long
  - Tweets are typically short, consisting of no more than 140 characters.
- Informal vs. Formal
  - Typos, abbreviations, phonetic substitutions, ungrammatical structures and use of emoticons.
  - Full of user generated words, urban words, E.g. kewl for cool!
- Conversational vs. Presentation
  - Tweets are conversational, hence individual tweet is often incomplete and needs the sequence to provide overall context.
  - Content is dynamic
  - Documents are more standalone

Tweets vs. Documents cont.

From user/distribution aspects:

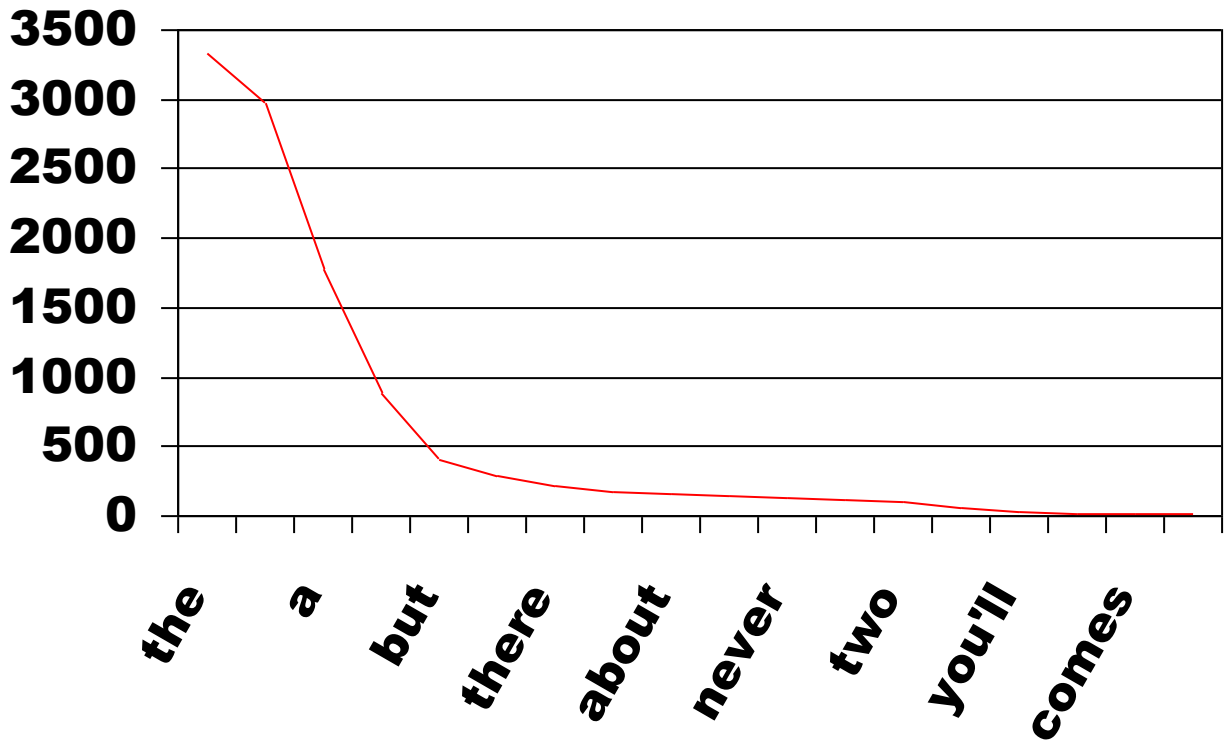
- Dynamic user community
  - Follower/followee relations
  - Various topical interests
  - Users come and go quickly
- Live data streams (key)
  - Data arrive continuously in a stream.
  - Real-time processing

Preprocessing for tweets

Similar to free-text document analysis

- Term extraction
  - Word segmentation for Chinese tweets
- Stopword removal
- Vocabulary normalization
- Term vector representation

Word Frequencies in Tom Sawyer

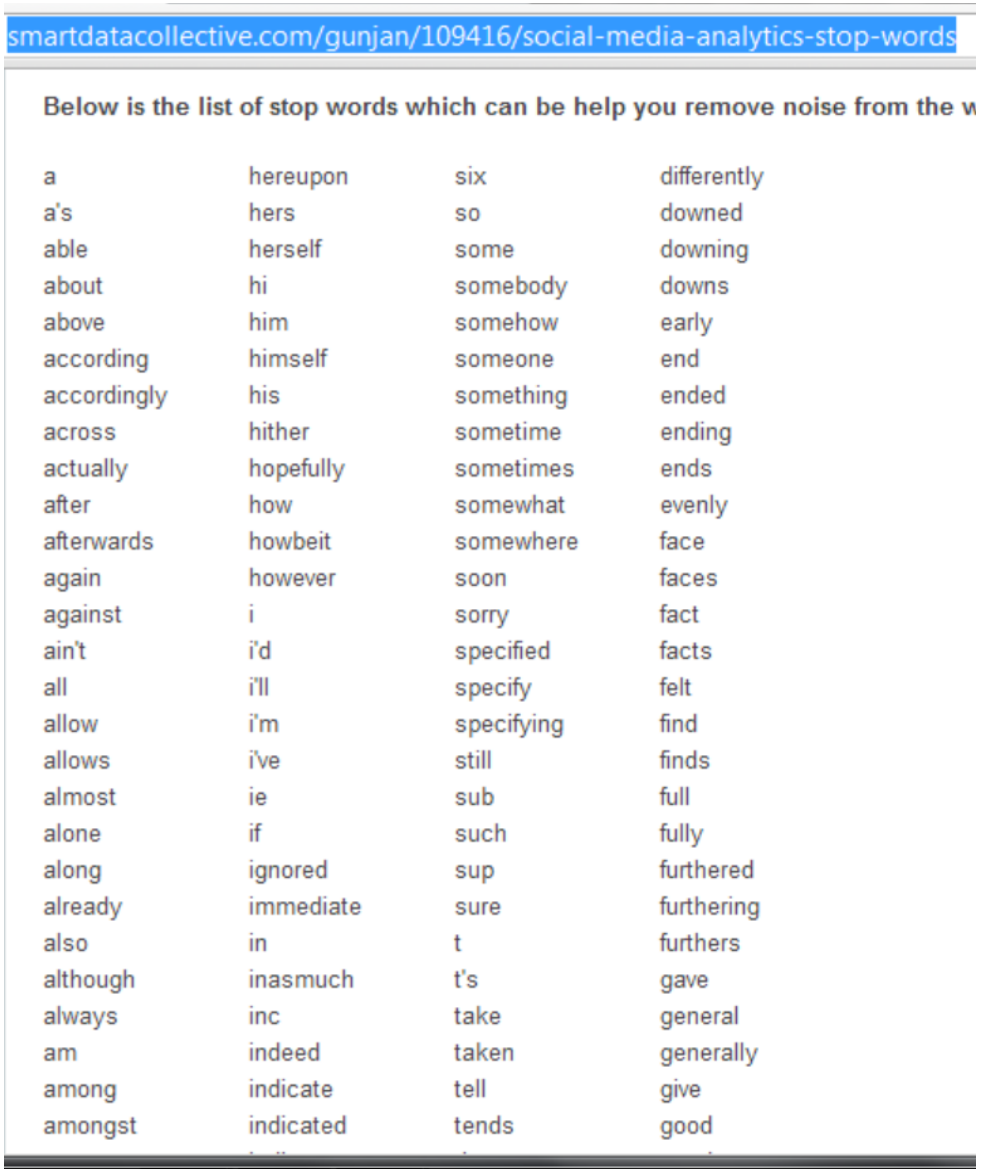


Stopword Removal

- Stopwords are words which are filtered out prior to, or after, processing of text.
- There is no one definite list of stop words which all systems use.
- Some systems specifically avoid removing them to support phrase search.

Examples of Stopword List

- Largely similar to normal text processing
- See: http://smartdatacollective.com/gunjan/109416/social-media-analytics-stop-words



smartdatacollective.com/gunjan/109416/social-media-analytics-stop-words

Below is the list of stop words which can be help you remove noise from the w

a	hereupon	six	differently
a's	hers	so	downed
able	herself	some	downing
about	hi	somebody	downs
above	him	somehow	early
according	himself	someone	end
accordingly	his	something	ended
across	hither	sometime	ending
actually	hopefully	sometimes	ends
after	how	somewhat	evenly
afterwards	howbeit	somewhere	face
again	however	soon	faces
against	i	sorry	fact
ain't	i'd	specified	facts
all	i'll	specify	felt
allow	i'm	specifying	find
allows	i've	still	finds
almost	ie	sub	full
alone	if	such	fully
along	ignored	sup	furthered
already	immediate	sure	furthering
also	in	t	furthers
although	inasmuch	t's	gave
always	inc	take	general
am	indeed	taken	generally
among	indicate	tell	give
amongst	indicated	tends	good

Resources for Stopword Removal

- Other Resources
- There is an in-built stopword list in NLTK made up of 2,400 stopwords for 11 languages (Porter et al) (see http://nltk.org/book/ch02.html)
- http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words
- http://snowball.tartarus.org/algorithms/english/stop.txt

Stemming

There are several types of stemming algorithms which differ in respect to performance and accuracy and how certain stemming obstacles are overcome.

A stemmer for ENGLISH, for example, should identify the STRING "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

Brut Force Stemming

- These stemmers employ a lookup table which contains relations between root forms and inflected forms. To stem a word, the table is queried to find a matching inflection. If a matching inflection is found, the associated root form is returned.
- Benefits.
  - Stemming error less.
  - User friendly.
- Problems
  - They lack elegance to converge to the result fast.
  - Time consuming.
  - Back end updating
  - Difficult to design.

.

Suffix Stemming

- Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" are stored which provide a path for the algorithm, given an input word form, to find its root form.
- Some examples of the rules include:
  - if the word ends in 'ed', remove the 'ed'
  - if the word ends in 'ing', remove the 'ing'
  - if the word ends in 'ly', remove the 'ly'
- Benefits:
  - Simple

Vocabulary Normalization

- Reduce variants of terms to standard form, like the role of stemming or thesaurus
- A substantial amount of tweets involve the use of informal expressions:

eg: se u 2morw!!!, cu tmr!!

-> See you tomorrow!

earthqu, eathquake, earthquakeee

-> standard form earthquake

b4 -> before

goooood -> good

- How many forms of variants are there??
  - Typos (goooooood)
  - Abbreviations (se, u, eartqu, …)
  - Phonetic substitutions (cu, b4, ..)
  - Can you think of any others??

Perform Vocabulary Normalization -1

- Cannot use stemming (as there are no regularities)
- The simplest is to detect lexical variants, and normalize lexical variants based on twitter dictionary.

- Resources
  eg: http://www.twittonary.com/
  http://www.csse.unimelb.edu.au/~tim/etc/emnlp2012-lexnorm.tgz

  - An English Social Media Normalization Lexicon [Han et al. 2012]
  - Contains about 40K (lexical variant, normalization) pairs automatically mined from 80 million English tweets from Sep 2010 to Jan 2011.
  - A crowd sourcing platform...

Perform Vocabulary Normalization -2

- Method
  - Given a tweet, we go through the dictionary and change any occurrences of informal expressions that are detected into their formal equivalent.
- With this approach, we can detect and correct a large proportion of informal expressions found within incoming tweets.

Overall Processing Pipeline

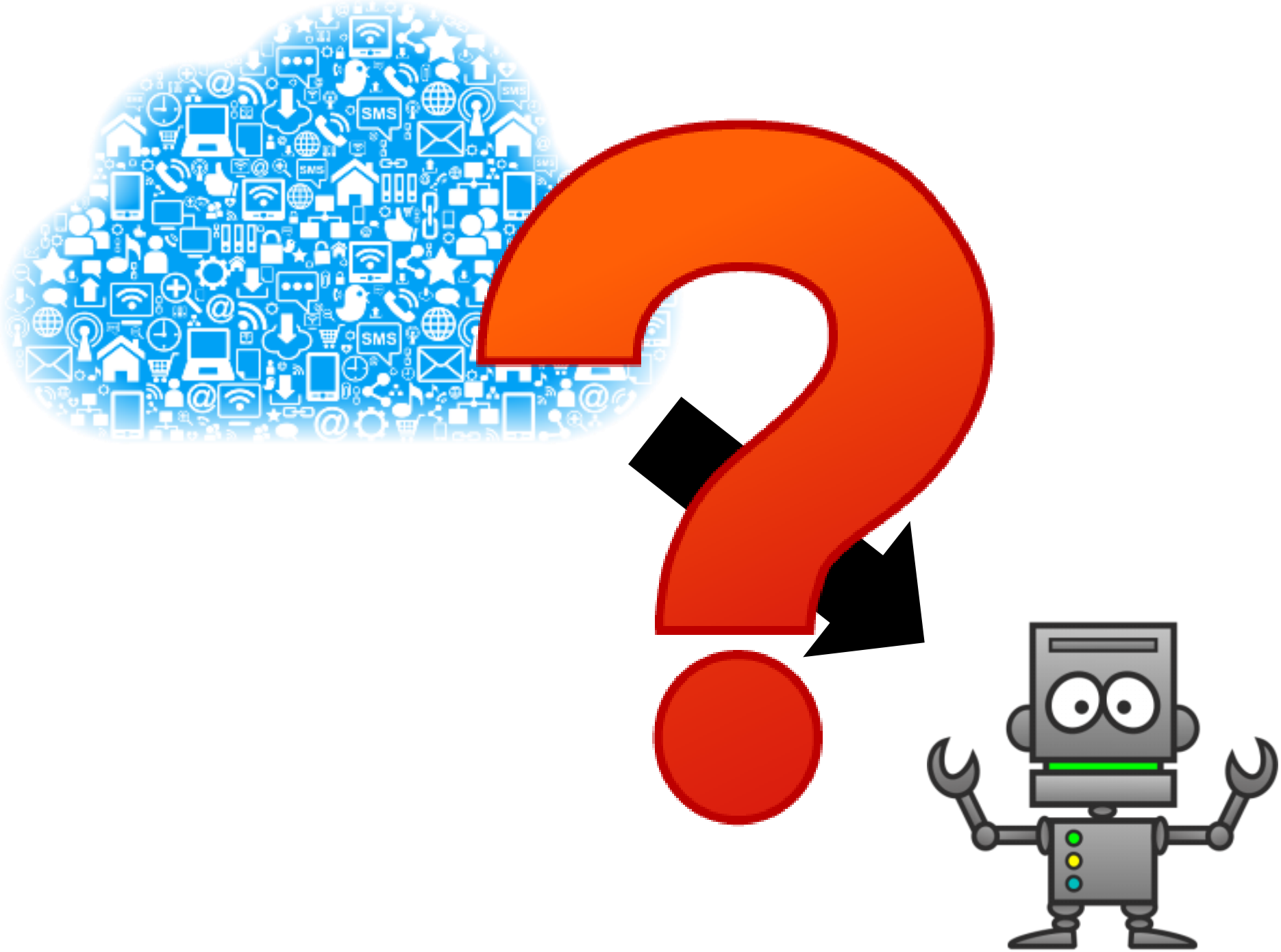
- The pre-processing module helps to correct for informal language usage to reduce errors that may be encountered downstream during feature extraction.
  - Language identification
  - Informal language normalization:

    to detect and standardize informal expressions found within incoming tweets.
  - Irrelevant text tokens filtering:

    to remove URLs, user mentions (*i.e. @username), retweet prefixes (i.e. RT followed by a sure name),* and non-alphabetical special characters.
  - Discard the tweet if the final length <= 3 characters

Contents

- What is Microblog
- Text Preprocessing
- Textual Data Representation
- Summary


SMS
SMS
SMS

N-Gram Models of Language

- **Use word sequences** of length n = 1… k, called n-grams
- Language Model (LM)
  - unigrams (n = 1), bigrams (n = 2), trigrams,…
- How do we obtain such data representations?
  - Very large corpora – **Why?**

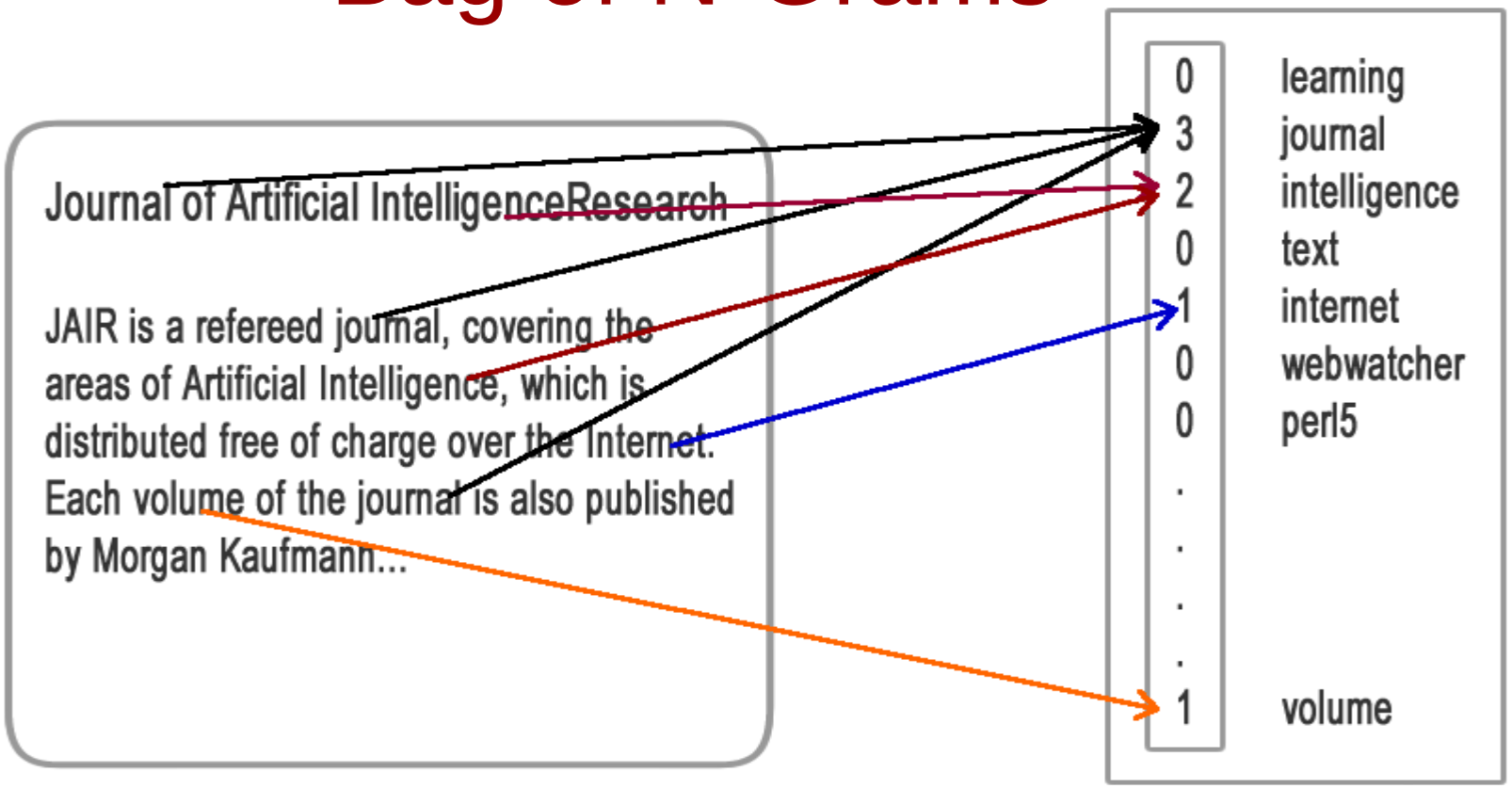
Simple N-Grams

- Assume a language has T words in its lexicon, how likely is word x to follow word y?
  - Simplest model of word probability: 1/T
  - Alternative 1: estimate likelihood of x occurring in new text based on its general frequency of occurrence estimated from a corpus (unigram probability)

    popcorn is more likely to occur than *unicorn*
  - Alternative 2: condition the likelihood of x occurring in the context of previous words (bigrams, trigrams,…)

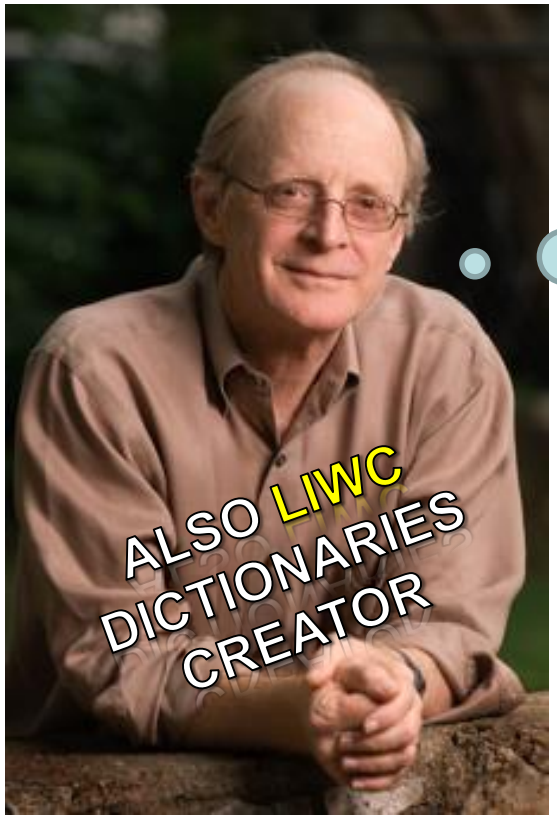
    mythical unicorn is more likely than *mythical popcorn*

Bag of N-Grams



- Count occurrences of each term (N-gram).
- Pick top N most frequent as a Bag of Terms (N-Grams): T = $\{t_1, t_2, t_3, \ldots, t_N\}$
- Each document D for user u is represented by normalized terms (N-grams) distribution among N most frequent terms (N-Grams): $D_u = \{\frac{t_{1u}}{N}, \frac{t_{2u}}{N}, \frac{t_{3u}}{N}, \ldots, \frac{t_{Nu}}{N}\}$

Words usage study for personality profiling



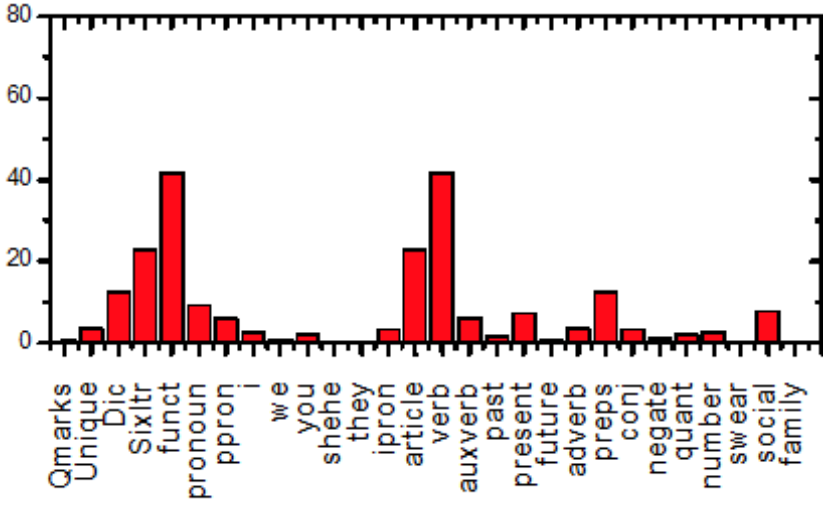
James W. Pennebaker

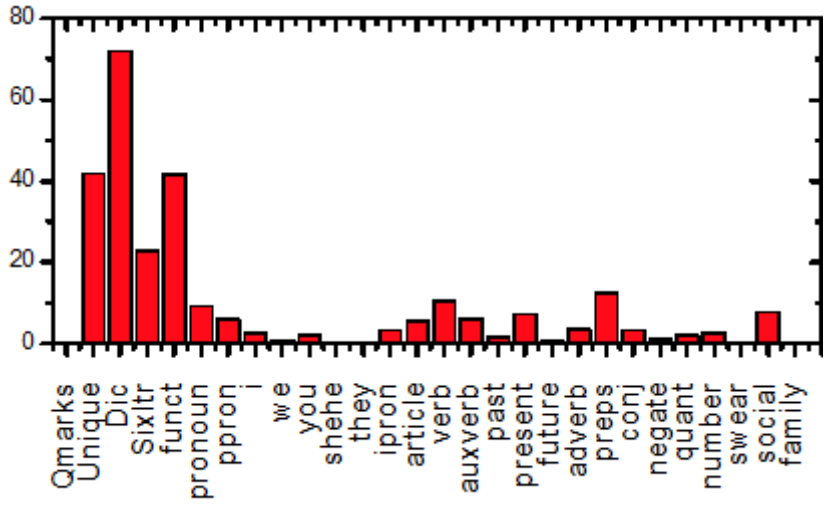
The smallest, most commonly used, most forgettable words serve as windows into our thoughts, emotions, and behaviors.

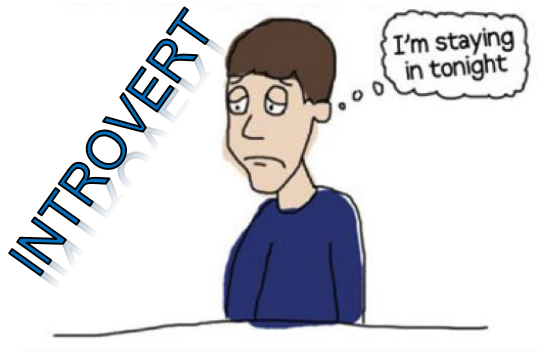
- Task – Word usage analysis* and correlation with personality
- Data – Various essays and questionnaires
- Approach – manual personality-related dictionaries construction
- Findings:
  - *Certain word usage statistics are good indicators for human personality profiling*

* Pennebaker, J. W. (2011). The secret life of pronouns.

LIWC

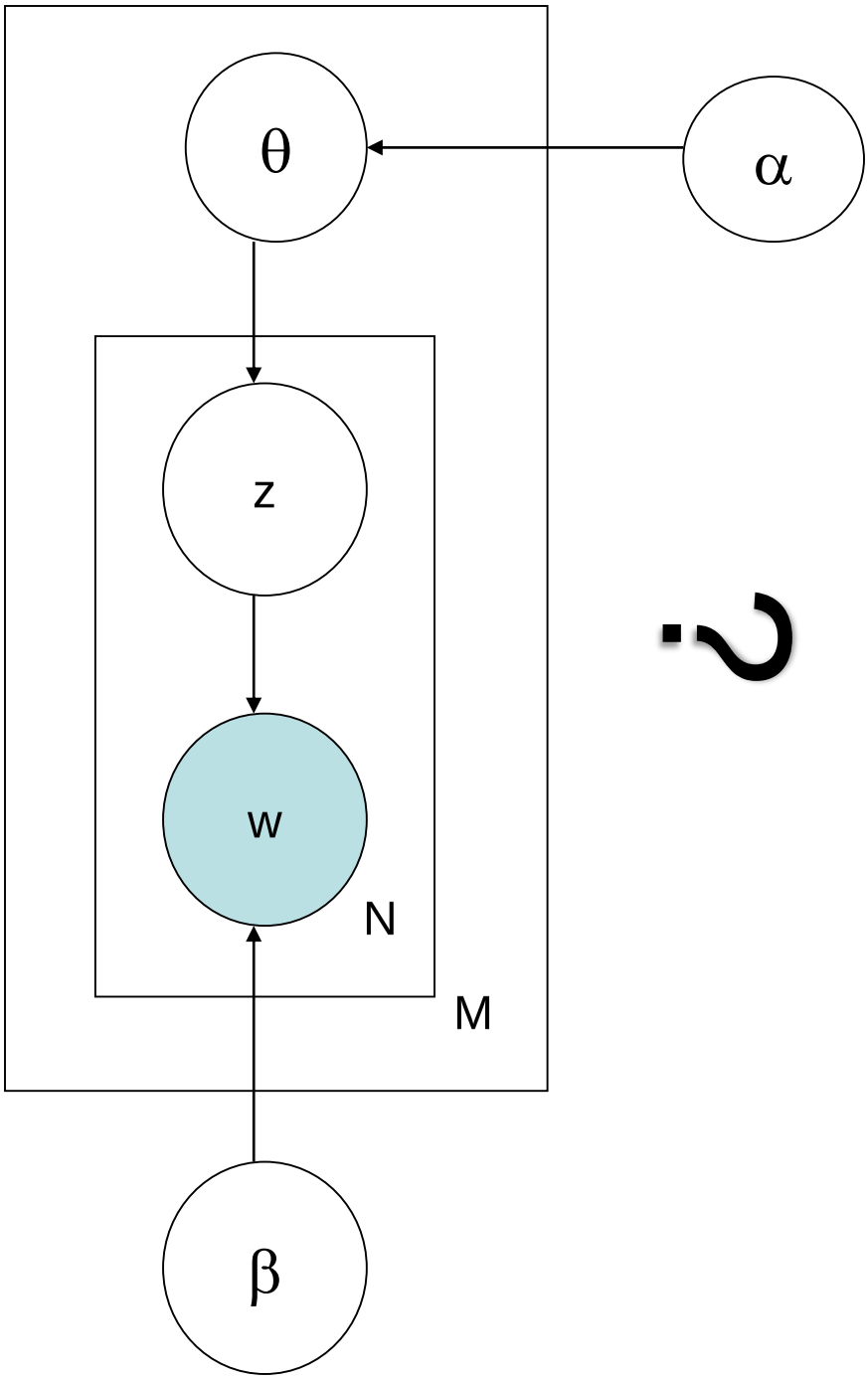






PERSONALITY

- Count occurrences of each LIWC category
- Each document D for user u is represented as a distribution among 74 LIWC categories: $D_u = \{\frac{LIWC_{1u}}{N}, \frac{LIWC_{2u}}{N}, \frac{LIWC_{3u}}{N}, \ldots, \frac{LIWC_{74u}}{N}\}$


θ
α
z
w
N
M
β

Topic Modeling -1

- Methods for automatically organizing, understanding, searching and summarizing large electronic archives.
  - Uncover hidden topical patterns in collections.
  - Annotate documents according to topics.
  - Using annotations to organize, summarize and search.
  - Widely poplar approach: Latent Dirichlet Allocation (LDA)*

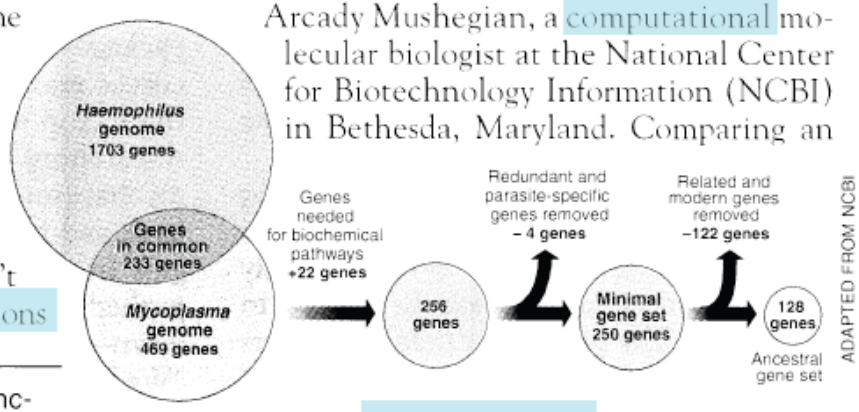
*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

Topic Modeling -2

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

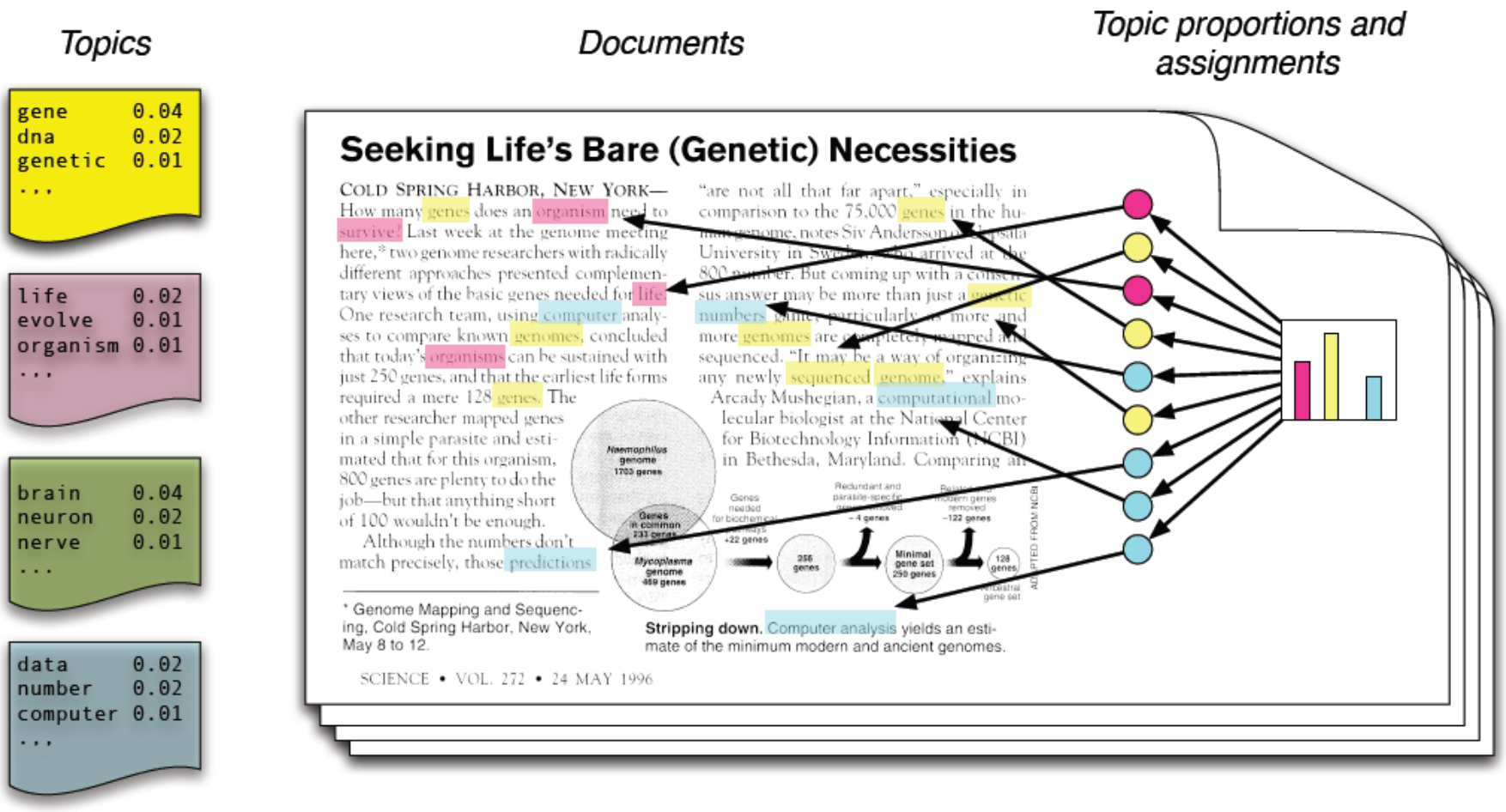
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Topic Modeling -3



Topic Modeling -4

- Only documents are observable (***All user's tweets are in one document for every user***).
- Infer underlying topic structure:
  - Topics that generated the documents.
  - For each document, distribution of topics.
  - For each word, which topic generated the word.

LDA – Data

- **Suppose we have the following set of sentences:**
  1. I like to eat broccoli and bananas.
  2. I ate a banana and spinach smoothie for breakfast.
  3. Chinchillas and kittens are cute.
  4. My sister adopted a kitten yesterday.
  5. Look at this cute hamster munching on a piece of broccoli.

Given these sentences and asked for 2 topics, LDA might produce something like:

- **Sentences 1 and 2**: 100% Topic A
- **Sentences 3 and 4**: 100% Topic B
- **Sentence 5**: 60% Topic A, 40% Topic B
- **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … *(at which point, we could interpret **topic A** to be about **food**)*
- **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … *(at which point, you could interpret **topic B** to be about **cute animals**)*

*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

LDA – Generative Process

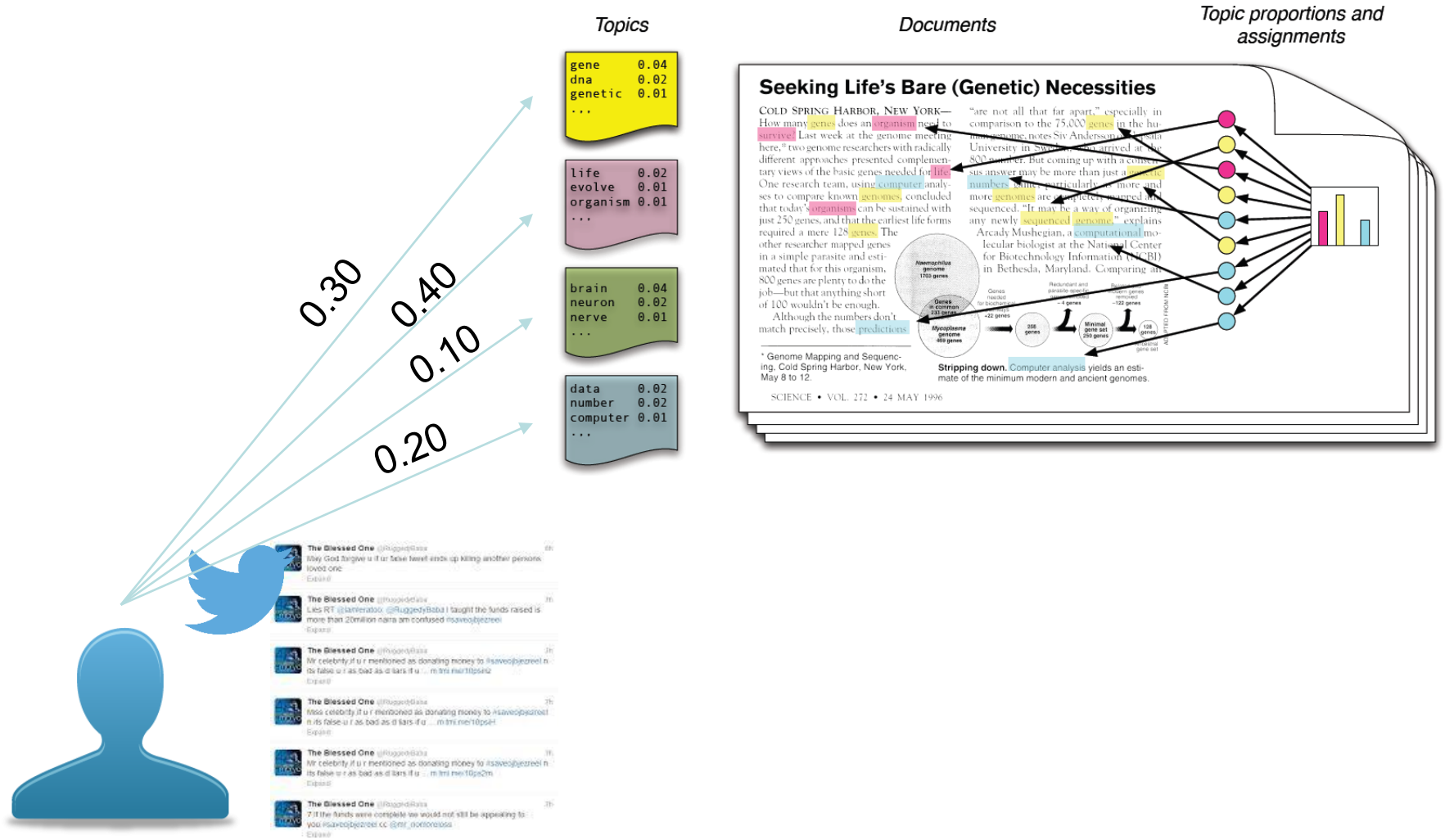
- LDA assumes that when writing each document, you:
  1. Decide on the number of words N the document will have (say, according to a Poisson distribution).
  2. Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). (*For example, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.*)
  3. Generate each word w_i in the document by:
     1. Picking a topic (according to the multinomial distribution that you sampled above); (*For example, we might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability*).
     2. Using the topic to generate the word itself (according to the topic's multinomial distribution). *(For example, if we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on.)*

*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

LDA – Learning Process

- LDA backtracks the generative process to recover topics from documents
- One way (collapsed Gibbs sampling) is the following:
  1. Go through each document, and randomly assign each word in the document to one of the K topics.
  2. Improve the assignment for each document by going through each word w_i in document d_j and for each topic t, compute :
     1. p(topic t | document d) = the proportion of words in document d that are assigned to topic t
     2. p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w.
     3. Reassign w a new topic, where we choose topic t with probability p(topic t | document d) * p(word w | topic t) (*according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability*).

*D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.



- Each document D (one user - one document) is represented as a distribution among N LDA topics: $D_u = \{ LDA_{1_u}, LDA_{2_u}, LDA_{3_u}, \ldots, LDA_{N_u} \}$

Behavioral Features

Feature	Description
Number of hash tags	Number of hash tags mentioned in message
Number of slang words	Number of slang words one use in his tweets. We calculate number of slang words / tweet and compute average slang usage
Number of URLs	Number of URL's one usually use in his/her tweets
Number of user mentions	Number of user mentions – may represent one's social activity
Number of repeated chars	Number of repeated characters in one tweets (e.g. nooooooooo, wahhhhhhh)
Number of emotion words	Number of words that are marked with not – neutral emotion score in Sentiment WordNet
Number of emoticons	Number of common emoticons from Wikipedia article
Average sentiment level	Module of average sentiment level of tweet obtained from Sentiment WordNet
Average sentiment score	Average sentiment level of tweet obtained from Sentiment WordNet
Number of misspellings	Number of misspellings fixed by Microsoft Word spell checker
Number Of Mistakes	Number of words that contains mistake but cannot be fixed by Microsoft Word spell checker
Number of rejected tweets	Number of tweets where 70% of words either not in English or cannot be fixed by Microsoft Word spell checker
Number of terms average	Average number of terms per / tweet

If we aim to do analysis on tweet level, what additional features can be used?

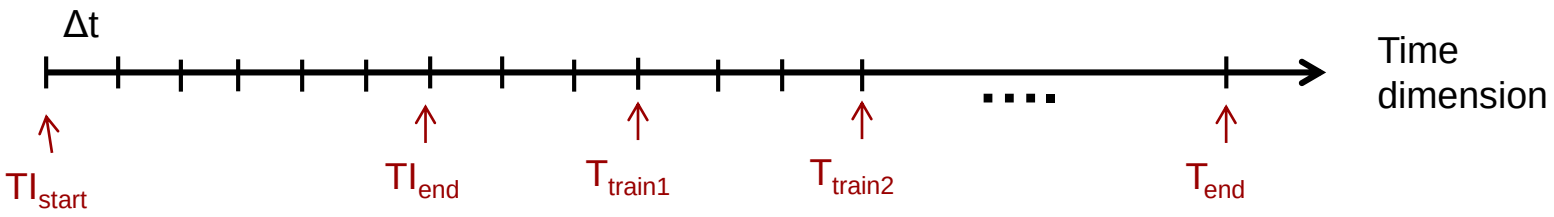
- Evolving Text Features
- Location Information
- User Social Relationships
- User Tweeting Tendencies Over Time

Evolving Text Features -1

- How to handle evolution of text?
- Two approaches:
  - Represent data based on latest set of text features
  - Assign lower weights to text terms that are not used recently

Evolving Text Features -2

- Given the timeline of tweet arrival:

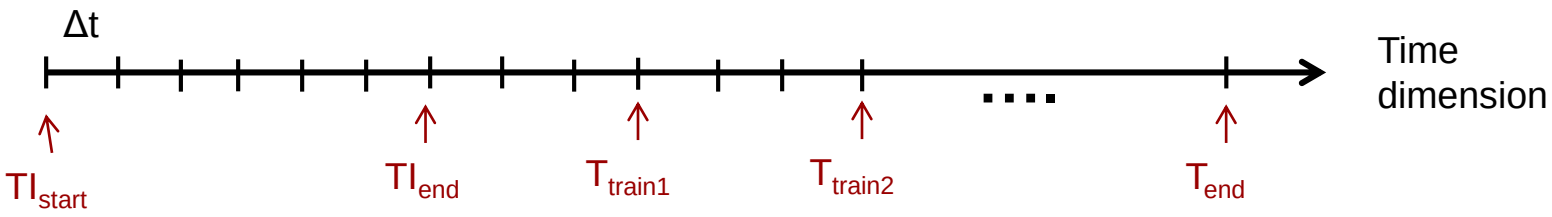


- Definitions:
  - $[T_{start}, T_{end}]$: the interval of the event
  - $[TI_{start}, TI_{end}]$: the initial window (IW)
  - $[TI_{end}, T_{train1}]$, $[T_{train1}, T_{train2}]$: dynamic training windows (DWs)

- Idea:
  - Incorporate text features extracted from IW and a latest DW as time advances
  - IW: ensures stable vocab and avoid topic drift
  - DW: ensures latest set of vocab is used.

Evolving Text Features -3

- The timeline of tweet arrival:



- Issues:
  - When to update IW?
  - What about older DWs? Should they be weighted less?
  - What is a good size of time interval Δt?
    should it be 6, 12, 24 or 48 hours??

Evolving Text Features -4

- Temporarily weighted text features:
  - Lexical and syntactic features have traditionally been important features for text processing.
  - May want to weight a recently used terms higher than those used some time ago
  - The governing equation for temporal term feature is:

$$c_i = \frac{\sum_j \theta^{(t_j - t)} |w_{ij}|}{\sum_i \sum_j \theta^{(t_j - t)} |w_{ij}|}$$

where $\theta > 1$ is the decay factor; t_j ($< t$) is the origin time of T_i; and w_{ij} is the term frequency of term t_j in Tweet T_i.

- The word feature set used at time t is:

$$L^t = \{(w_1, c_1), (w_2, c_2), \ldots, (w_m, c_m)\}$$ → Known as $\mathbf{F_c}$

Location Features -1

- For the case of “NUS”, what is the best way to differentiate “National University of Singapore” vs. “National Union of Students”?

  Answer: Use location if it can be found.

- Previous studies showed that location plays a big part in the contents of tweets. This correlation is intuitive as people will often discuss or talk about events happening around them.
  - Given the “NUS” example, a tweet containing this acronym from a user based in Singapore will likely be referring to the “National University of Singapore”, whereas the same tweet by a user based in UK is more likely to be about “National Union of Students”.

Location Features -2

- Three key sources of location information.
- User Profile:
  - The location info stated in users' profiles, or the time zone they reside
  - 66% of users included valid geo location info at city level
- Geo-location
  - More tweets come with geo-tagged info now, though the percentage is still low in 2015 (about 1% only)
  - Can map geo tag to a geographical country using OpenHeatMap
- Inferring Geo Location from text
  - Given appropriate textual evidence, geo location can be inferred to about 70% accuracy with geographical location accurate to about 10Km.

Location Features -3

- The location feature set used is (known as $\mathbf{F_d}$):
  - Location Difference:

    whether users' profile location is the same with that of desired topic?
  - Time zone Difference:

    whether users' profile time zone is the same with that of desired topic?
  - Geo-tagged Difference:

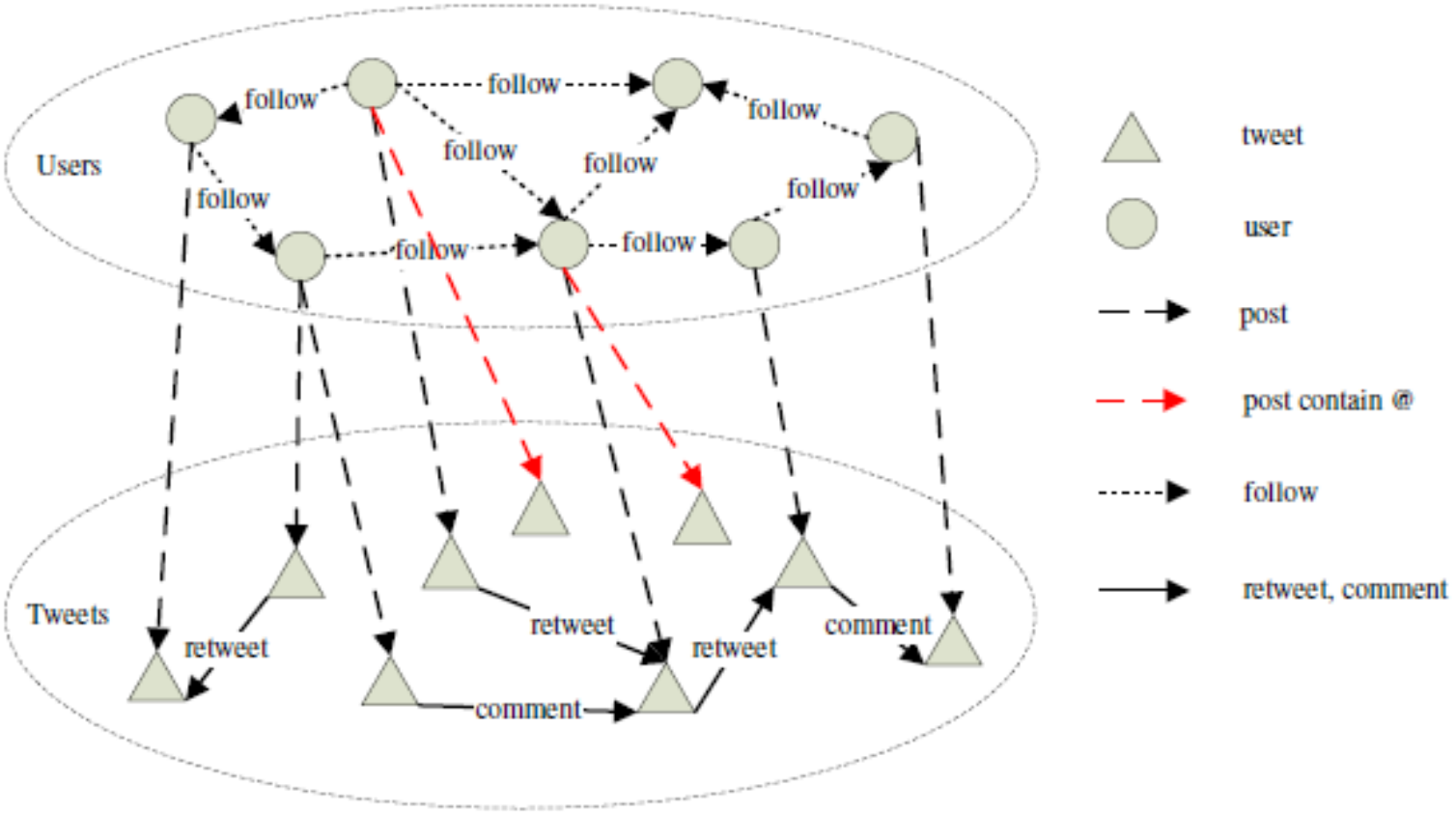
    whether location of geo-tagged tweets (at country level) is the same with that of desired topic?

User Social Relationships -1

- We note that social relationships include both explicit social relationships and implicit social relationships.
- Explicit social relationships refer to formal ways user accounts can be associated together on a microblog service.
  - For example in the case of Twitter, an explicit social relationship exists between two users if at least one of the users ***"follows"*** *the other.*
- Implicit social relationships refer to when users interact with one another on a microblog service via:
  - Interactions: comments , re-tweets, reply etc.
  - Others implicit links may be established based on similar profile, similar topics od interests etc.

User Social Relationships -2

- Allow us to build up an overview of users who may potentially share similar interests or are related via a common affiliation or activity.



User Social Relationships -3

- Tweet relevance can be inferred from social relations
- Leads to social feature set: $\mathbf{F_{s1}}$:
  - Interact from relevant tweet

    whether current tweet is a re-tweet or comment from a relevant tweet
  - Interact from irrelevant tweet

    whether current tweet is a re-tweet or comment from an irrelevant tweet
  - Follow relevant user

    whether the user of current tweet follows a relevant user account
  - Follow irrelevant user

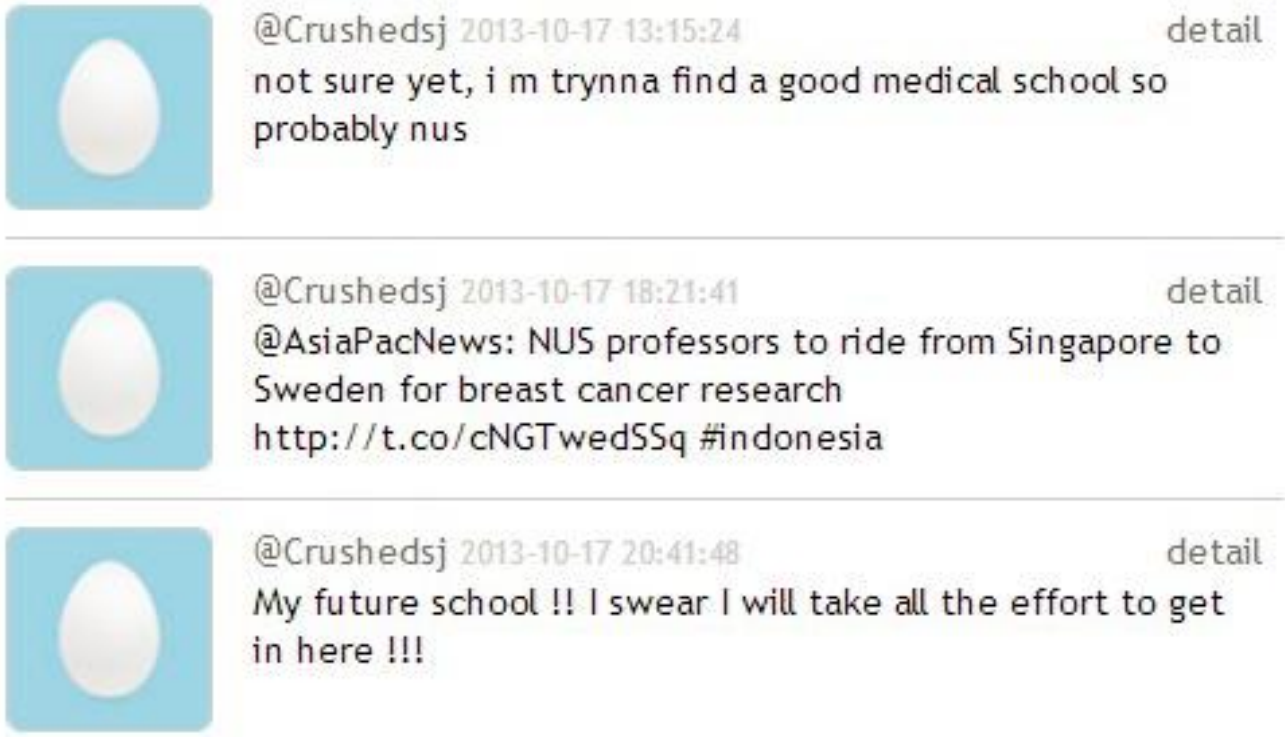
    whether the user of current tweet follows an irrelevant user account

User Social Relationships -4

- For organizations or important accounts, there are known accounts, that frequent tweet about the entity:
  - For example NUS has more than 10 twitter accounts
  - These accounts offer relevant tweets and relevant user groups
  - Based on our studies, 80% of users related to an known account are within 2 edges of a social graph away from the relevant known accounts
- Based on known accounts, we can define further social features as: $\mathbf{F_{s2}}$:
  - Distance to relevant known account
  - Comment on relevant known account
  - Referred to relevant known account
  - Distance to irrelevant known account
  - Comment on irrelevant known account
  - Referred to irrelevant known account

User Tweeting Tendencies Over Time -1

- Another Observation: Tweets refer to a relevant event may or may not contain similar keywords
- We propose to analyze past tweets a user made to infer the relevance of current tweets
- Example
  - 3 tweets sent within 24 hours of each other
  - First 2 refer to "NUS", while the last tweet no
  - Based on earlier tweets, we can infer that last tweet is relevant to NUS



User Tweeting Tendencies Over Time -2

- Important empirical observations:
  - About 70-80% of tweets do not contain references to organization names
  - Up to 17% and 29% of users from Twitter and Weibo respectively make more than one tweet about the same event within the same day
- User Tweeting Tendency Feature set, $\mathbf{F_t}$:
  - Immediate relevancy

    whether last tweet by same user within time span dT is relevant
  - Trend relevancy

    whether majority of tweets by user in time span dT is relevant

Contents

- What is Microblog
- Text Preprocessing
- Textual Data Representation
- Summary

Summary

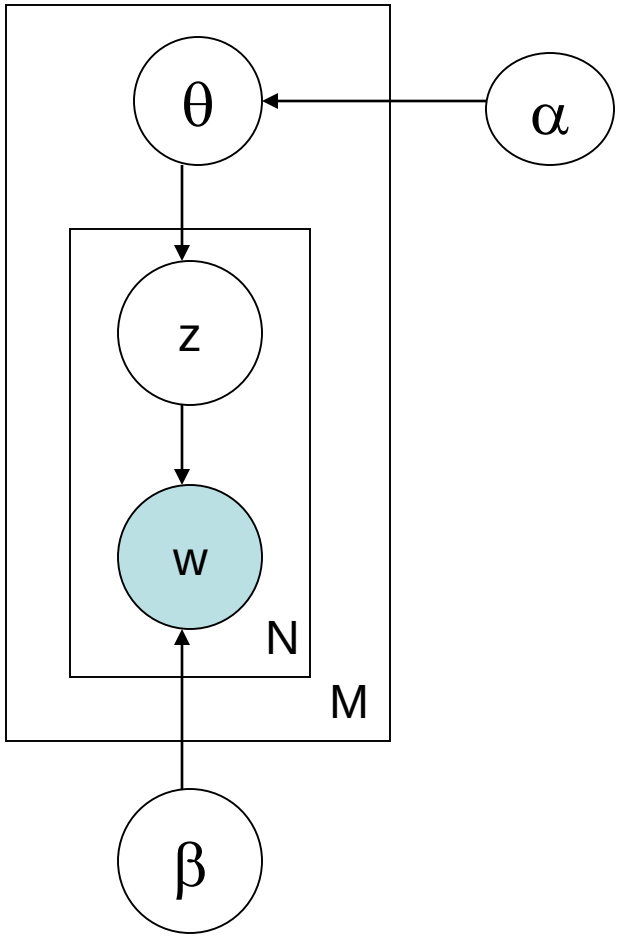
- Microblogs are shorter, and much more noisy as compared to other text sources (Blogs, Wikipedia)
- Textual Data always need to be pre-processed:
  - Stop words removal
  - Vocabulary normalization
- Different Feature Types Could be extracted:
  - Bag of N-grams (Unigrams, or words)
  - Linguistic features (i.e. LIWC)
  - Latent Topics (i.e. LDA)
  - Behavioral Features (i.e. mistakes, sentiment, activity level)
  - Relations:
    - Spatial (location)
    - Temporal (terms evolution over time)
    - Social (social graph)

Next Lesson

- **Location and Image Data Processing**

Backup slides

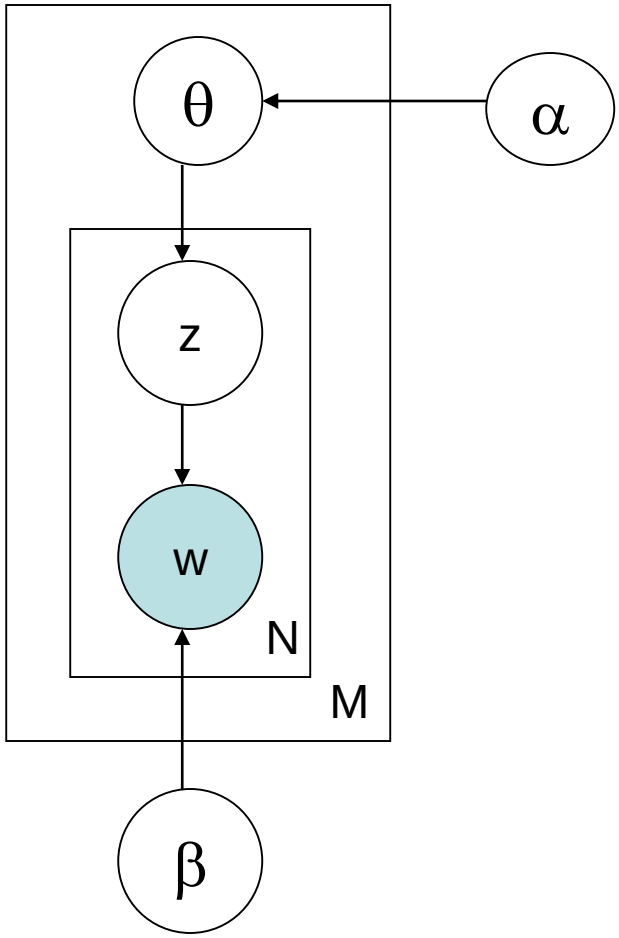
Topic Modeling -backup
Latent Dirichlet Allocation (LDA)



- **for each** document d = 1,⋯,M
  - Generate $\theta_d \sim Dir(\cdot \mid \alpha)$
  - **for each (word)** position n = 1,⋯, N_d
    - Generate $z_n \sim Mult(\cdot \mid \theta_d)$
    - Generate $w_n \sim Mult(\cdot \mid \beta_{z_n})$

- α is the parameter of the Dirichlet prior on the per-document topic distributions,
- β is the parameter of the Dirichlet prior on the per-topic word distribution,
- θ_d is the topic distribution for document *d*,
- β_{z_n} is the word distribution for topic *k*,
- z_n is the topic for the *n*th word in document *d*
- w_n is the specific word.

Topic Modeling -backup
Learning LDA



- From a collection of documents M, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
- Use posterior expectation to perform different tasks.