POIR 613: Computational Social Science

Pablo Barberá

School of International Relations
University of Southern California
`pablobarbera.com`

Course website:

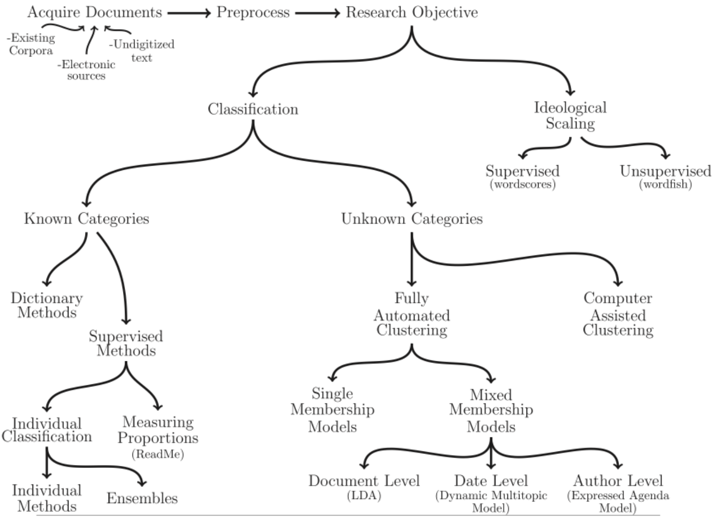
pablobarbera.com/POIR613/

Today

1. Project
   - Peer feedback was due on Monday
   - Next milestone: 5-page summary that includes some data analysis by November 4th
2. Topic models
3. Solutions to challenge 6
4. Additional methods to compare documents

Topic models

Overview of text as data methods



Outline

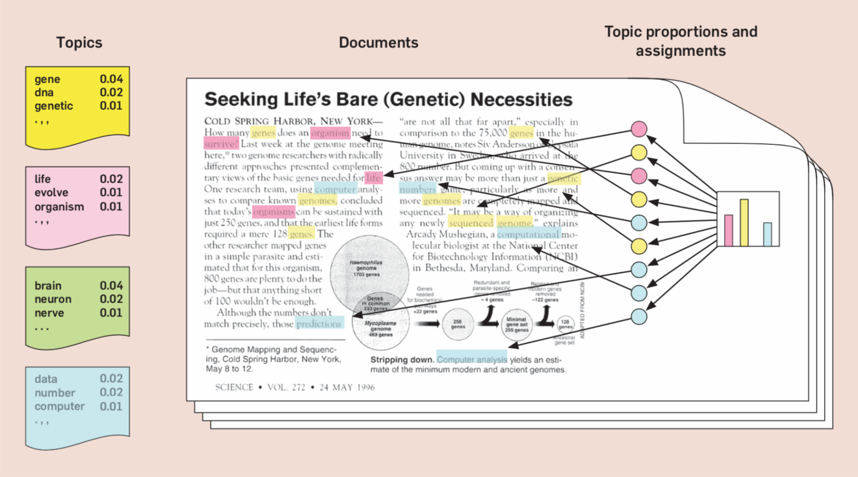
- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Validating the output of topic models
- Examples
- Choosing the number of topics
- Extensions of LDA

Topic Models

- Topic models are algorithms for discovering the main “themes” in an unstructured corpus
- Can be used to organize the collection according to the discovered themes
- Requires no prior information, training set, or human annotation – only a decision on K (number of topics)
- Most common: Latent Dirichlet Allocation (LDA) – Bayesian mixture model for discrete data where topics are assumed to be uncorrelated
- LDA provides a generative model that describes how the documents in a dataset were created
  - Each of the K *topics* is a distribution over a fixed vocabulary
  - Each document is a collection of words, generated according to a multinomial distribution, one for each of K topics

Latent Dirichlet Allocation

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Validating the output of topic models
- Examples
- Choosing the number of topics
- Extensions of LDA

Latent Dirichlet Allocation

- Document = random mixture over latent topics
- Topic = distribution over n-grams

Probabilistic model with 3 steps:

1. Choose $\theta_i \sim \mathrm{Dirichlet}(\alpha)$
2. Choose $\beta_k \sim \mathrm{Dirichlet}(\delta)$
3. For each word in document i:
   - Choose a topic $z_m \sim \mathrm{Multinomial}(\theta_i)$
   - Choose a word $w_{im} \sim \mathrm{Multinomial}(\beta_{i,k=z_m})$

where:

α=parameter of Dirichlet prior on distribution of topics over docs.

θ_i=topic distribution for document i

δ=parameter of Dirichlet prior on distribution of words over topics

β_k=word distribution for topic k

Latent Dirichlet Allocation

Key parameters:

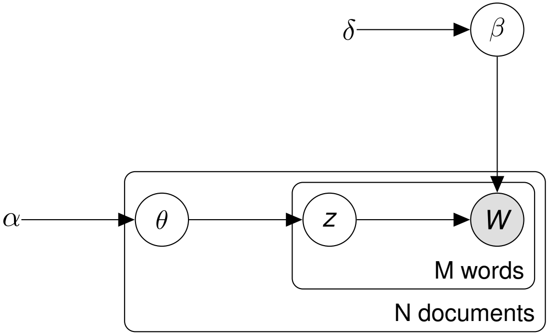
1. θ = matrix of dimensions N documents by K topics where θ_{ik} corresponds to the probability that document i belongs to topic k; i.e. assuming $K = 5$:

	T1	T2	T3	T4	T5
Document 1	0.15	0.15	0.05	0.10	0.55
Document 2	0.80	0.02	0.02	0.10	0.06
...					
Document N	0.01	0.01	0.96	0.01	0.01

2. β = matrix of dimensions K topics by M words where β_{km} corresponds to the probability that word m belongs to topic k; i.e. assuming $M = 6$:

	W1	W2	W3	W4	W5	W6
Topic 1	0.40	0.05	0.05	0.10	0.10	0.30
Topic 2	0.10	0.10	0.10	0.50	0.10	0.10
...						
Topic k	0.05	0.60	0.10	0.05	0.10	0.10

Plate notation



$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m
$\theta = N \times K$ matrix where θ_{ik} indicates prob(topic=k) for document i

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Validating the output of topic models
- Examples
- Choosing the number of topics
- Extensions of LDA

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?
3. Predictive validity
   - Does variation in topic usage correspond with expected events?
4. Hypothesis validity
   - Can topic variation be used effectively to test substantive hypotheses?

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Validating the output of topic models
- Examples
- Choosing the number of topics
- Extensions of LDA

Example: open-ended survey responses

Bauer, Barberá *et al*, *Political Behavior*, 2016.

- Data: General Social Survey (2008) in Germany
- Responses to questions: *Would you please tell me what you associate with the term "left"? and would you please tell me what you associate with the term "right"?*
- Open-ended questions minimize priming and potential interviewer effects
- Sparse Additive Generative model instead of LDA (more coherent topics for short text)
- $K = 4$ topics for each question

Example: open-ended survey responses

Table 1: Top scoring words associated with each topic, and English translations)

Left topic 1: **Parties** (proportion = .26, average lr-scale value = 5.38)
linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks
the left, spd, party, the left, pds, politics, communists, parties, greens, punks

Left topic 2: **Ideologies** (proportion = .26, average lr-scale value = 5.36)
kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei
communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling

Left topic 3: **Values** (proportion = .24, average lr-scale value = 4.06)
soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung
social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights

Left topic 4: **Policies** (proportion = .24, average lr-scale value =4.89)
sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten
social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent

Right topic 1: **Ideologies** (proportion = .27, average lr-scale value = 5.00)
konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative
conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives

Right topic 2: **Parties** (proportion = .25, average lr-scale value = 5.26)
npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikalen
npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicalists

Right topic 3: **Xenophobia** (proportion = .25, average lr-scale value = 4.55)
ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus
xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism

Right topic 4: **Right-wing extremists** (proportion = .23, average lr-scale value = 4.90)
nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale
nazis, neonazis, right-wing radicalists, rightists, radicals, radicalism, party, xenophobia, rich, national

Note: "proportion" indicates the average estimated probability that any given response is assigned to a topic. "average lr-scale value" is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: open-ended survey responses

Fig. 6: Left-right scale means for different subsamples of associations with **left** (dashed = sample mean, bars = 95% Cis)

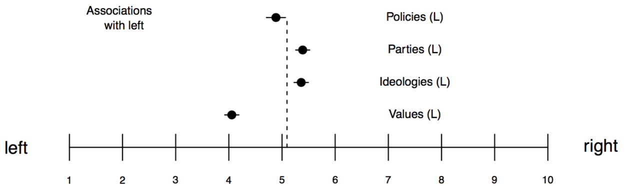
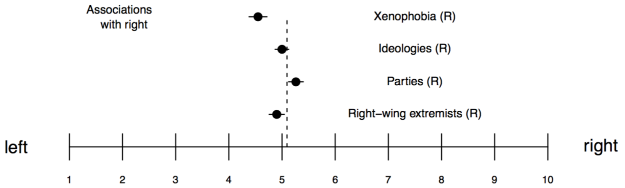


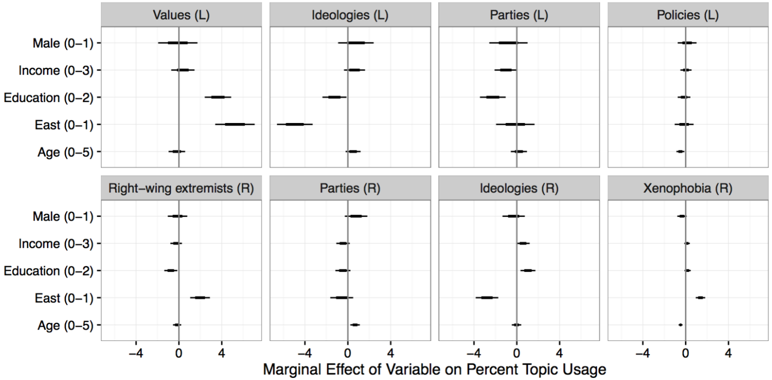
Fig. 7: Left-right scale means for different subsamples of associations with **right** (dashed = sample mean, bars = 95% Cis)



Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: open-ended survey responses

Fig. 9: Systematic relationship between associations with "left" and "right" and characteristics of respondents



Note: Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated "right" with political parties.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: topics in US legislators' tweets

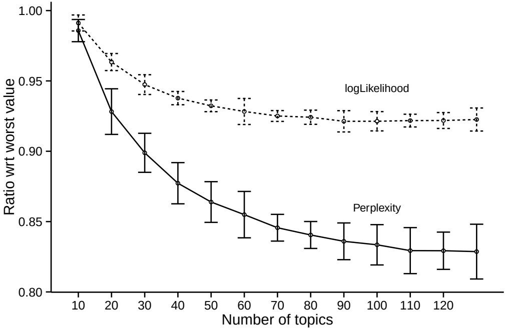
- Data: 651,116 tweets sent by US legislators from January 2013 to December 2014.
- 2,920 documents = 730 days $\times$ 2 chambers $\times$ 2 parties
- Why aggregating? Applications that aggregate by author or day outperform tweet-level analyses (Hong and Davidson, 2010)
- $K = 100$ topics (more on this later)
- Validation: http://j.mp/lda-congress-demo

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Validating the output of topic models
- Examples
- Choosing the number of topics
- Extensions of LDA

Choosing the number of topics

- Choosing K is "one of the most difficult questions in unsupervised learning" (Grimmer and Stewart, 2013, p.19)
- One approach is to decide based on cross-validated model fit



- **BUT**: "there is often a negative relationship between the best-fitting model and the substantive information provided".
- GS propose to choose K based on "substantive fit."

Model evaluation using "perplexity"

- can compute a likelihood for "held-out" data
- perplexity: can be computed as (using VEM):

$$\text{perplexity}(w) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\}$$

- lower perplexity score indicates better performance

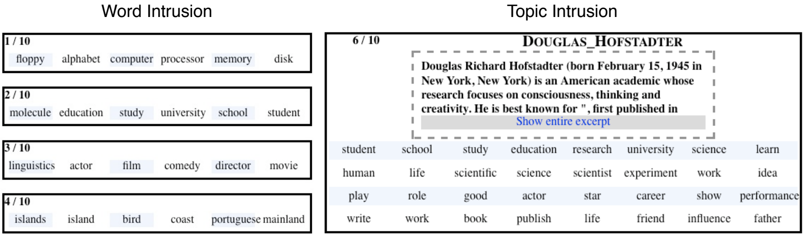
Evaluating model performance: human judgment

(Chang, Jonathan et al. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in neural information processing systems*.)

Uses human evaluation of:

- whether a topic has (human-identifiable) semantic coherence: word intrusion, asking subjects to identify a spurious word inserted into a topic
- whether the association between a document and a topic makes sense: topic intrusion, asking subjects to identify a topic that was not associated with the document by the model

Example



- conclusions: the quality measures from human benchmarking were negatively correlated with traditional quantitative diagnostic measures!

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Validating the output of topic models
- Examples
- Choosing the number of topics
- Extensions of LDA

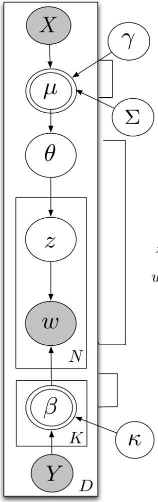
Extensions of LDA

1. Structural topic model (Roberts et al, 2014, AJPS)
2. Dynamic topic model (Blei and Lafferty, 2006, ICML; Quinn et al, 2010, AJPS)
3. Hierarchical topic model (Griffiths and Tenembaun, 2004, NIPS; Grimmer, 2010, PA)

Why?

- Substantive reasons: incorporate specific elements of DGP into estimation
- Statistical reasons: structure can lead to better topics.

Structural topic model



Topic Prevalence:

$$\begin{aligned}\mu_{d,k} &= X_d \gamma_k \\ \gamma_k &\sim \mathcal{N}(0, \sigma_k^2) \\ \sigma_k^2 &\sim \text{Gamma}(s^\gamma, r^\gamma)\end{aligned}$$

Language Model:

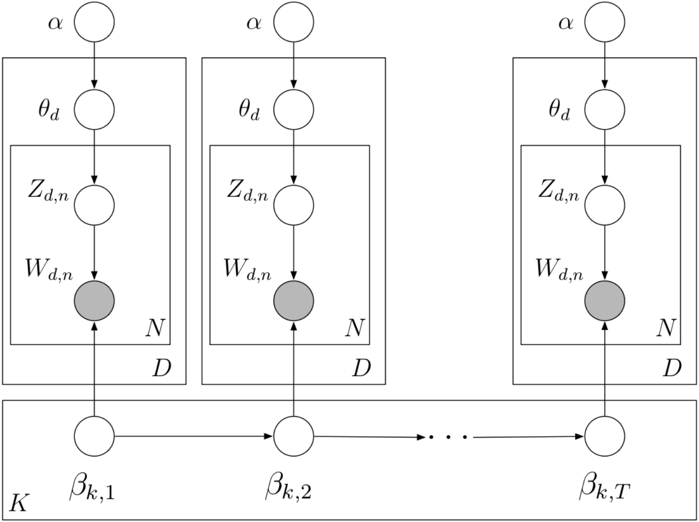
$$\begin{aligned}\theta_d &\sim \text{LogisticNormal}(\mu_d, \Sigma) \\ z_{d,n} &\sim \text{Mult}(\theta_d) \\ w_{d,n} &\sim \text{Mult}(\beta_d^{k=z_{d,n}})\end{aligned}$$

Topical Content:

$$\begin{aligned}\beta_{d,v}^k &\propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k}) \\ \kappa_v^{y,k} &\sim \text{Laplace}(0, \tau_v^{y,k}) \\ \tau_v^{y,k} &\sim \text{Gamma}(s^\kappa, r^\kappa)\end{aligned}$$

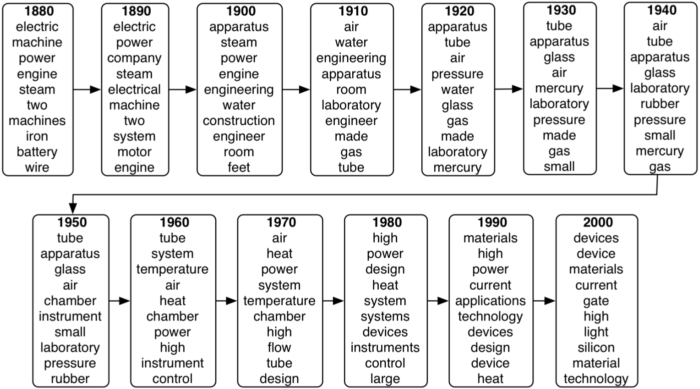
- **Prevalence**: Prior on the mixture over topics is now document-specific, and can be a function of covariates (documents with similar covariates will tend to be about the same topics)
- **Content**: distribution over words is now document-specific and can be a function of covariates (documents with similar covariates will tend to use similar words to refer to the same topic)

Dynamic topic model



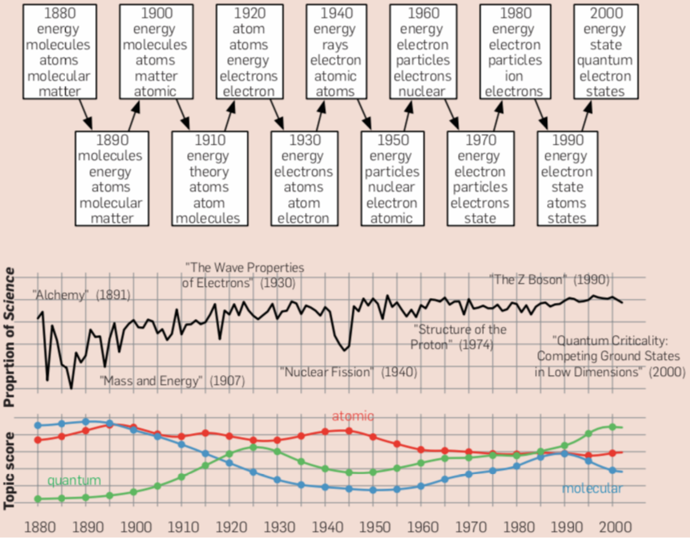
Source: Blei, “Modeling Science”

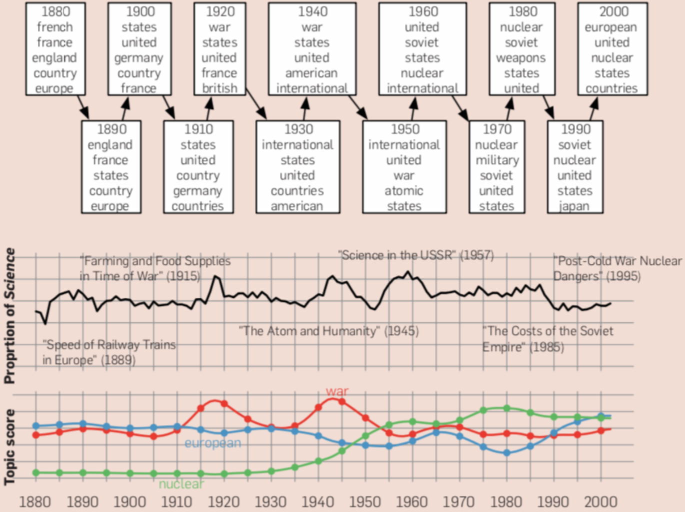
Dynamic topic model



Source: Blei, “Modeling Science”

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.





1880 french france england country europe
1890 england france states country europe
1900 states united germany country france
1910 states united country germany countries
1920 war states united france british
1930 international states united countries american
1940 war states united american international
1950 international united war atomic states
1960 united soviet states nuclear international
1970 nuclear military soviet united states
1980 nuclear soviet weapons states united
1990 soviet nuclear united states japan
2000 european united nuclear states countries
Proprtion of Science
"Farming and Food Supplies in Time of War" (1915)
"Science in the USSR" (1957)
"Post-Cold War Nuclear Dangers" (1995)
"Speed of Railway Trains in Europe" (1889)
"The Atom and Humanity" (1945)
"The Costs of the Soviet Empire" (1985)
Topic score
war
european
nuclear
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

Comparing documents

- Describing a single document
  - Lexical diversity
  - Readability
- Comparing documents
  - Similarity metrics: cosine, Euclidean, edit distance
  - Clustering methods: *k*-means clustering

Quantities for describing a document

Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Word (relative) frequency counts or proportions of words

Lexical diversity (At its simplest) involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens

Readability statistics Use a combination of syllables and sentence length to indicate “readability” in terms of complexity

Lexical Diversity

- Basic measure is the TTR: Type-to-Token ratio
- Problem: This is very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions
- Another problem: length may relate to the introduction of additional subjects, which will also increase richness

Lexical Diversity: Alternatives to TTRs

TTR $\frac{\text{total types}}{\text{total tokens}}$

Guiraud $\frac{\text{total types}}{\sqrt{\text{total tokens}}}$

S Summer's Index: $\frac{\log(\log(\text{total types}))}{log(log(\text{total tokens}))}$

MATTR the Moving-Average Type-Token Ratio (Covington and McFall, 2010) calculates TTRs for a moving window of tokens from first to last token. MATTR is the mean of the TTRs of each window.

Readability

- Use a combination of syllables and sentence length to indicate “readability” in terms of complexity
- Common in educational research, but could also be used to describe textual complexity
- No natural scale, so most are calibrated in terms of some interpretable metric

Flesch-Kincaid readability index

- F-K is a modification of the original Flesch Reading Ease Index:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

  Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

- Flesch-Kincaid rescales to the US educational grade levels (1–12):

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

- Describing a single document
  - Lexical diversity
  - Readability
- Comparing documents
  - Similarity metrics: cosine, Euclidean, edit distance
  - Clustering methods: k-means clustering

Comparing documents

- The idea is that (weighted) features form a vector for each document, and that these vectors can be judged using metrics of similarity
- A document's vector for us is simply (for us) the row of the document-feature matrix
- The question is: how do we measure distance or similarity between the vector representation of two (or more) different documents?

Euclidean distance

Between document A and B where j indexes their features, where y_{ij} is the value for feature j of document i

- Euclidean distance is based on the Pythagorean theorem
- Formula

$$\sqrt{\sum_{j=1}^{j}(y_{Aj} - y_{Bj})^2} \tag{1}$$

- In vector notation:

$$\|\mathbf{y}_A - \mathbf{y}_B\| \tag{2}$$

- Can be performed for any number of features J (where J is the number of columns in of the dfm, same as the number of feature types in the corpus)

Cosine similarity

- Cosine distance is based on the size of the angle between the vectors
- Formula

$$\frac{\mathbf{y}_A \cdot \mathbf{y}_B}{\|\mathbf{y}_A\| \|\mathbf{y}_B\|} \tag{3}$$

- The $\cdot$ operator is the dot product, or $\sum_j y_{Aj} y_{Bj}$
- The $\|\mathbf{y}_A\|$ is the vector norm of the (vector of) features vector $\mathbf{y}$ for document A, such that $\|\mathbf{y}_A\| = \sqrt{\sum_j y_{Aj}^2}$
- Nice property: independent of document length, because it deals only with the angle of the vectors
- Ranges from -1.0 to 1.0 for term frequencies

Edit distances

- Edit distance refers to the number of operations required to transform one string into another for strings of equal length
- Common edit distance: the Levenshtein distance
- Example: the Levenshtein distance between ”kitten” and ”sitting” is 3
  - kitten $\rightarrow$ sitten (substitution of ”s” for ”k”)
  - sitten $\rightarrow$ sittin (substitution of ”i” for ”e”)
  - sittin $\rightarrow$ sitting (insertion of ”g” at the end).

Outline

- Describing a single document
  - Lexical diversity
  - Readability
- Comparing documents
  - Similarity metrics: cosine, Euclidean, edit distance
  - Clustering methods: k-means clustering

The idea of ”clusters”

- Essentially: groups of items such that inside a cluster they are very similar to each other, but very different from those outside the cluster
- “unsupervised classification”: cluster is not to relate features to classes or latent traits, but rather to estimate membership of distinct groups
- groups are given labels through post-estimation interpretation of their elements
- typically used when we do not and never will know the “true” class labels
- issues:
  - how many clusters?
  - which features to include?
  - how to compute distance is arbitrary

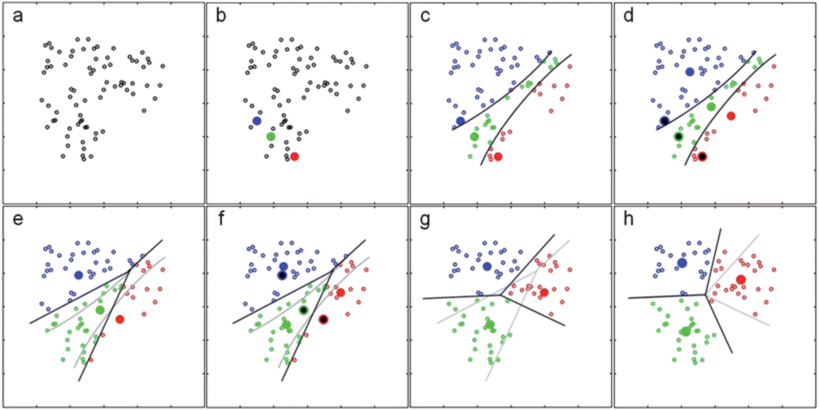
k-means clustering

- Essence: assign each item to one of *k* clusters, where the goal is to minimized within-cluster difference and maximize between-cluster differences
- Uses random starting positions and iterates until stable
- *k*-means clustering treats feature values as coordinates in a multi-dimensional space
- Advantages
  - simplicity
  - highly flexible
  - efficient
- Disadvantages
  - no fixed rules for determining *k*
  - uses an element of randomness for starting values

algorithm details

1. Choose starting values
   - assign random positions to *k* starting values that will serve as the "cluster centres", known as "centroids" ; or,
   - assign each feature randomly to one of *k* classes
2. assign each item to the class of the centroid that is "closest"
   - Euclidean distance is most common
   - any others may also be used (Manhattan, Minkowski, Mahalanobis, etc.)
   - (assumes feature vectors are normalized within document)
3. update: recompute the cluster centroids as the mean value of the points assigned to that cluster
4. repeat reassignment of points and updating centroids
5. repeat 2–4 until some stopping condition is satisfied
   - e.g. when no items are reclassified following update of centroids

k-means clustering illustrated



choosing the appropriate number of clusters

- very often based on prior information about the number of categories sought
  - for example, you need to cluster people in a class into a fixed number of (like-minded) tutorial groups
- a (rough!) guideline: set $k = \sqrt{N/2}$ where N is the number of items to be classified
  - usually too big: setting k to large values will improve within-cluster similarity, but risks *overfitting*
- "elbow plots": fit multiple clusters with different k values, and choose k beyond which are diminishing gains