POIR 613: Computational Social Science

Pablo Barberá

University of Southern California
`pablobarbera.com`

Course website:
pablobarbera.com/POIR613/

Comparing documents

- Describing a single document
  - Lexical diversity
  - Readability
- Comparing documents
  - Similarity metrics: cosine, Euclidean, edit distance
  - Keyness statistics
  - 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: $\frac{\log(\log(\text{total types}))}{\mathit{log}(\mathit{log}(\text{total tokens}))}$

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

- Based on the 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
  - Keyness statistics
  - 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
  - Keyness statistics
  - Clustering methods: k-means clustering

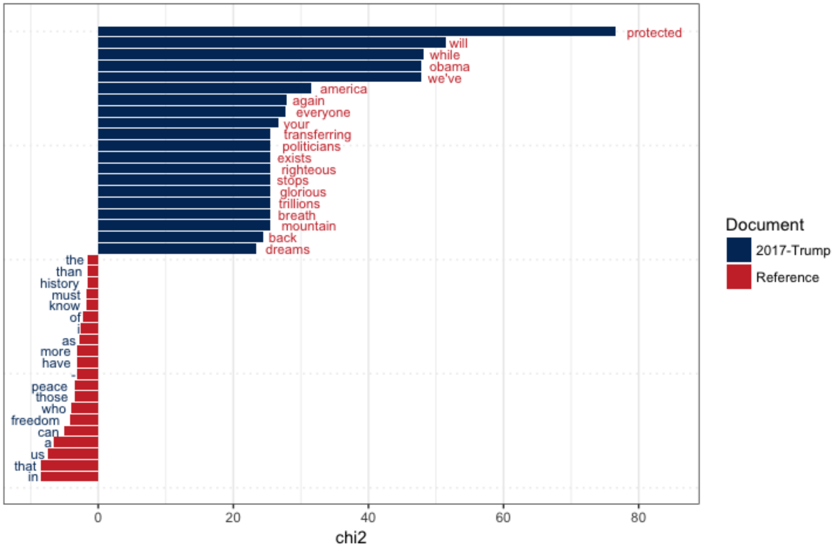
Keyness statistics

	Target	~ Target	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

χ^2 Pearson's χ^2 statistic, computed as:

$$\sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

where m_{ij} represents the cell frequency expected according to independence; i.e. $m_{ij} = n \times (\frac{n_i}{n} \times \frac{n_j}{n})$


protected
will
while
obama
we've
america
again
everyone
your
transferring
politicians
exists
righteous
stops
glorious
trillions
breath
mountain
back
dreams
the
than
history
must
know
of
i
as
more
have
-
peace
those
who
freedom
can
a
us
that
in
0
20
40
60
80
chi2
Document
2017-Trump
Reference

Outline

- Describing a single document
  - Lexical diversity
  - Readability
- Comparing documents
  - Similarity metrics: cosine, Euclidean, edit distance
  - Keyness statistics
  - 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

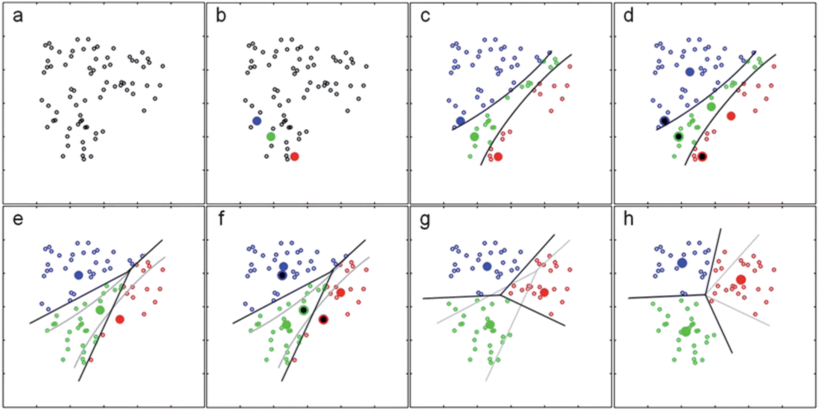
k-means clustering

- 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