Latent Semantic Indexing / Latent Semantic Analysis

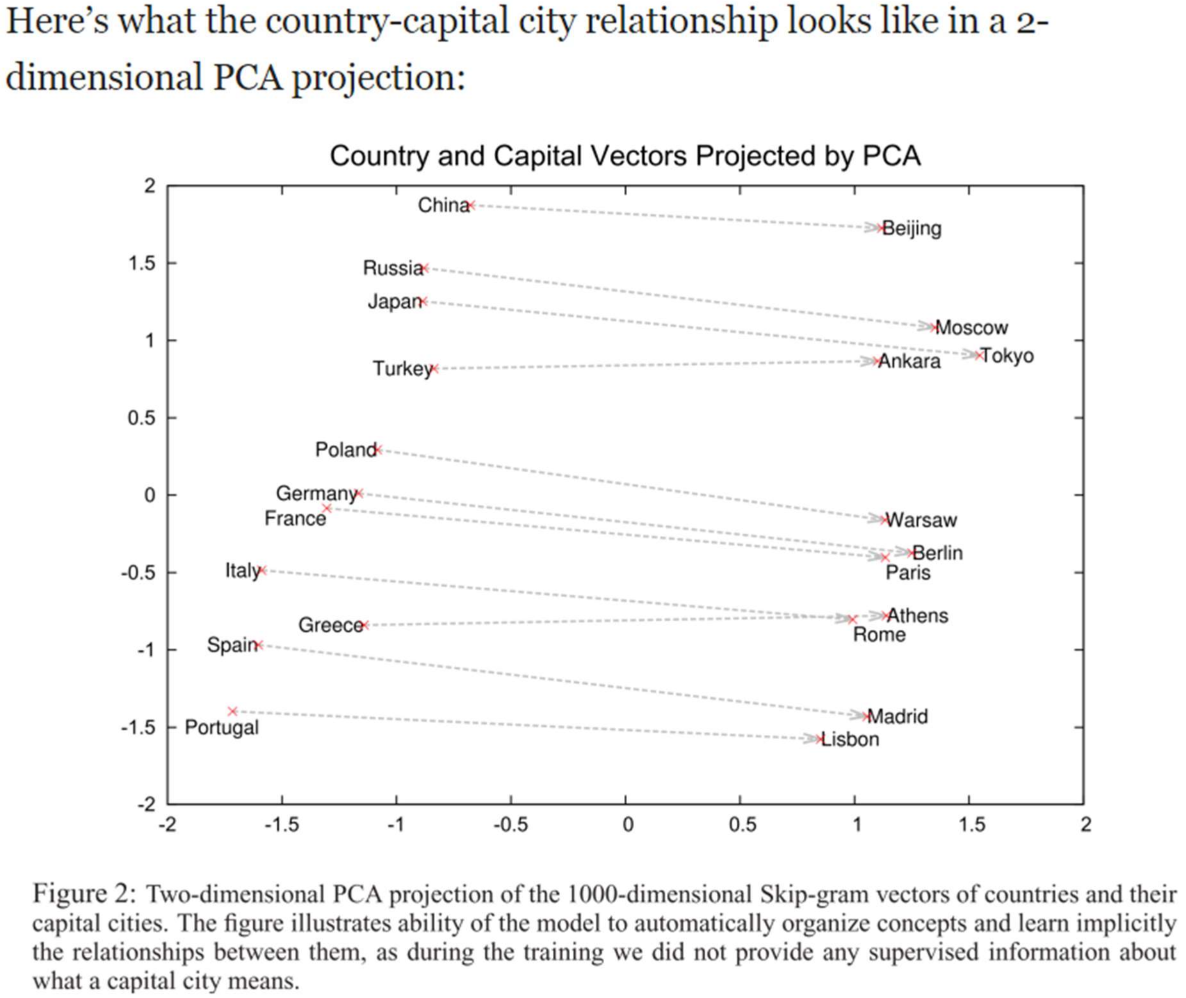
(with Introduction to Naive Bayes Text Classification)
Nicole Erich, Olivia Hong, Michael Lundy
August 10, 2016 - 2nd Half

Overview

For this module, we continued to explore latent semantic indexing to find similar documents. We created a more sophisticated vector-space that addresses some of the problems from the first section, like synonymy and polysemy.

Applications

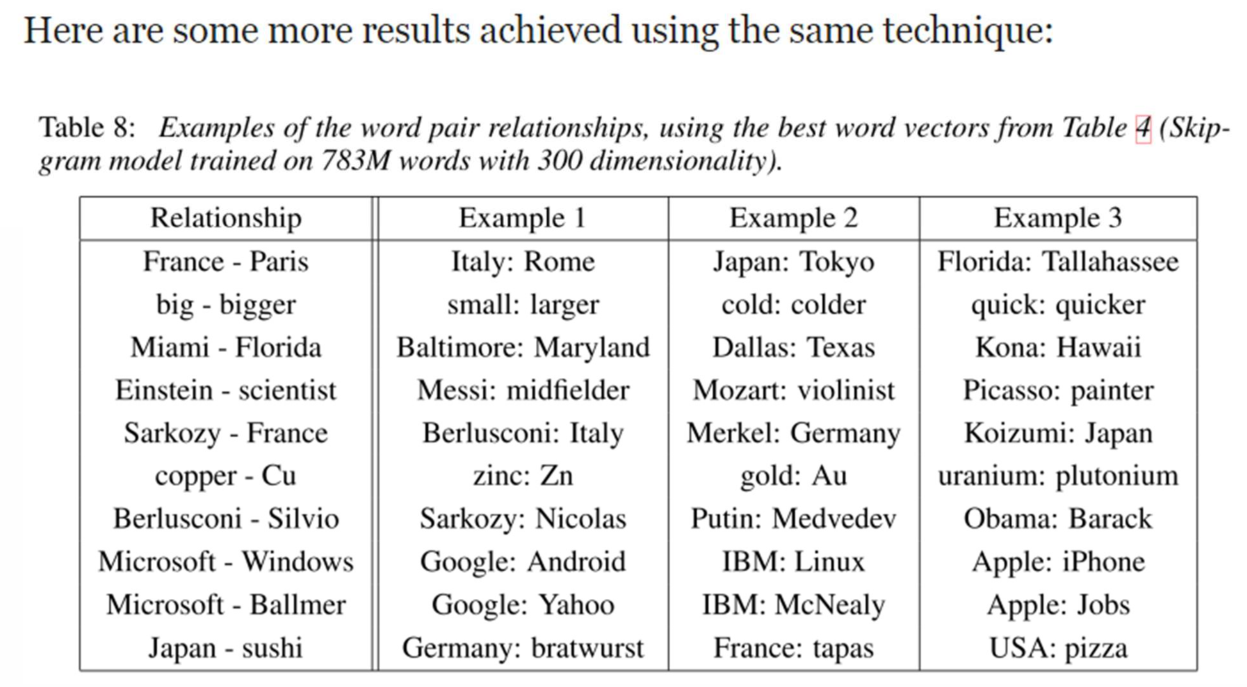
We'll look at a couple examples of vector-space from this blog post:
https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/



Here's what the country-capital city relationship looks like in a 2-dimensional PCA projection:

Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

The Skip-gram is a more complicated vector-space model. We can see where Portugal and Lisbon live in the vector space. Portugal is to Lisbon as Spain is to Madrid (analogy), according to their positions. The vector from Portugal to Lisbon is the Portugal vector (line from origin to Portugal) minus the Lisbon vector.



Here are some more results achieved using the same technique:

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

The model was trained on a vast quantity of data. As the quote goes, “More data almost always beats a smarter model.” We can see here it is capable of finding relationships that are similar in analogy, such as “Japan: sushi” and “Germany: bratwurst.”

Review the Math

1. Run PCA (Principal Components Analysis) on the original Z matrix of documents
2. From this model, obtain estimates of $v_1 \dots v_k$
   a. Where vectors represent the loadings / contexts
3. Using existing documents $z_1, z_2, \dots, z_n$ (where n is the number of documents)
   Calculate the score of each document i for the loading / context ℓ:

$$s_{i\ell} = z_i \cdot v_\ell$$

4. Result: for each doc_i, there is a vector s_i that contains the scores for doc_i on each of the principal components

$$s_i = s_{i1}, s_{i2}, \dots, s_{ik}$$

Key Point:
We now have a matrix S of documents (N) by contexts (K). Remember that in the original Z matrix, we had documents (N) by words (D).

K (the number of contexts) << D (the number of words)

$$S = \begin{pmatrix} s_1 & \rightarrow \\ s_2 & \rightarrow \\ . & \rightarrow \\ . & \rightarrow \\ . & \rightarrow \\ s_n & \rightarrow \end{pmatrix}$$

S is a new vector space:

- Rows are each document
- Each column represents a different loading/context
- The values are the score each document has for the given loading/context
- In this vector space, documents that score highly in similar contexts will be close in distance, no longer dependent on having the same words

Key Point:

- Before, similarity between documents was defined as the cosine of the angle between z_1 and z_2
- Now, similarity between documents is defined as the cosine of the angle between s_1 and s_2 (where s is the vector of scores of the document for each principal component)

Example: Which documents are similar to a new document?

1. Obtain q = the new query document as a D-dimensional vector of all the TF-IDF scores for each word

$$q = (q_1, \ldots, q_D)$$

2. Score q against all loadings / contexts (this is a K-dimensional vector that "lives" in the same space as matrix S)

$$t = (q \cdot v_1, q \cdot v_2, \ldots q \cdot v_k)$$

3. For each document i, compute:

$$\frac{s_i \cdot t}{|s_i|\cdot|t|}$$

So why is it called Latent Semantic indexing?

- Think of an index in a biology textbook as an example
- Finding all the pages that match a certain word or phrase will get most of the information, but an index needs to point to all pages that match certain content
- Latent semantic indexing will help to match the content, rather than just the word

R Script

Now a different vector-space representation: LSI/LSA.

```
lsi_art = prcomp(art_stories_DTM_TFIDF, scale.=FALSE)
```

Highly and lowly loaded words on the first PC (negative ones are the important ones).

```
head(sort(lsi_art$rotation[,1], decreasing=FALSE), 20)
```

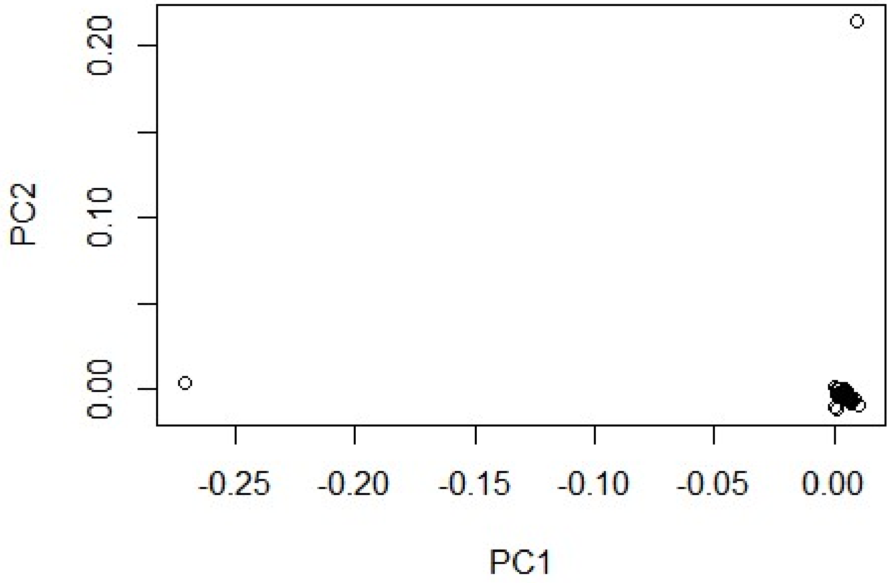
```
##     memorial       statue       queens      service metropolitan 
##   -0.3738331   -0.3680079   -0.3095011   -0.2242164   -0.1933623 
##      dispute        rabin        slain     undraped      yitzhak 
##   -0.1869166   -0.1869166   -0.1869166   -0.1869166   -0.1869166 
##         zeus neighborhood    misstated      article        leave 
##   -0.1869166   -0.1547418   -0.1547322   -0.1546395   -0.1542200 
##      israeli       report         lent       museum     decision 
##   -0.1540540   -0.1359199   -0.1358381   -0.1276675   -0.1224634
```

```
tail(sort(lsi_art$rotation[,1], decreasing=FALSE), 20)
```

```
##         her   recalling           n    athletes   botanical       drugs 
## 0.005780361 0.005781801 0.005815762 0.005829363 0.005856685 0.005856685 
##       stems           i         ill        will          he        weir 
## 0.005856685 0.005900520 0.006250449 0.006357104 0.006680037 0.007007723 
##       tolle          ms        said    patterns   tomaselli          mr 
## 0.007184291 0.007267999 0.007906044 0.008409741 0.008785027 0.009174935 
##       calif       teams 
## 0.009330823 0.010417415
```

Scores on the first two PCs. s1 and s2 from the S matrix are the axes here (The scores of each document for PC1 vs. the scores for each docuent for PC2). That is, we are plotting the documents to see how similar they are in the first two contexts.

```
plot(lsi_art$x[,1:2])
```



There are two clear outliers on the plot. We identify these two points using the identify function. These two outliers are weird stories about a zeus statue and a bridge tournament on the art page that don't relate to the other stories on the page in any meaningful ways. This is seen in these documents word, which gives them weighted scores far off from the other documents in regard to these principal components.

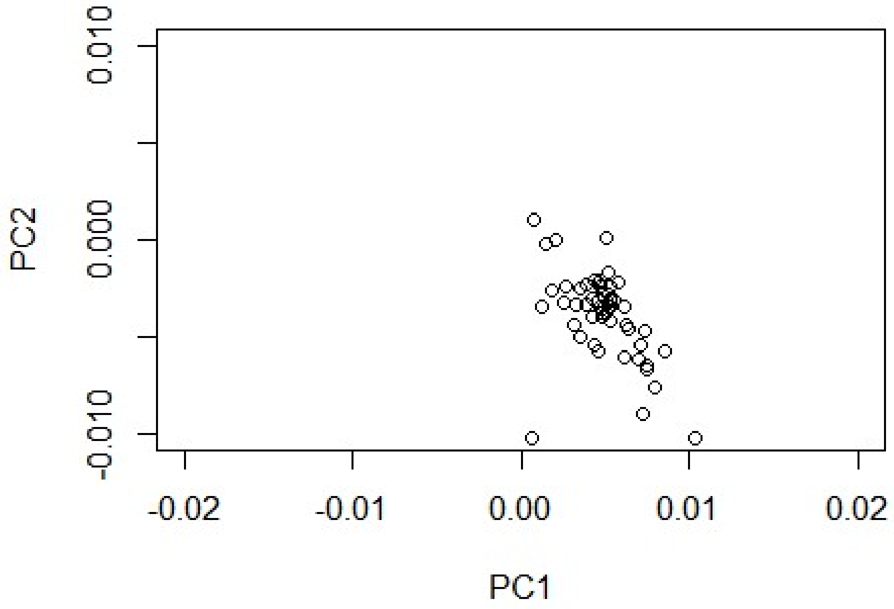
```
#identify(lsi_art$x[,1:2], n=2)
art_stories[[16]]

##  [1] "an"           "article"      "in"           "the"
##  [5] "neighborhood" "report"       "last"         "sunday"
##  [9] "about"        "a"            "dispute"      "over"
## [13] "whether"      "a"            "naked"        "statue"
## [17] "of"           "zeus"         "at"           "the"
## [21] "queens"       "museum"       "of"           "art"
## [25] "should"       "be"           "displayed"    "during"
## [29] "a"            "memorial"     "service"      "for"
## [33] "the"          "slain"        "israeli"      "prime"
## [37] "minister"     "yitzhak"      "rabin"        "misstated"
## [41] "the"          "role"         "of"           "the"
## [45] "metropolitan" "museum"       "the"          "metropolitan"
## [49] "lent"         "the"          "statue"       "to"
## [53] "the"          "queens"       "museum"       "but"
## [57] "was"          "not"          "involved"     "in"
## [61] "the"          "decision"     "to"           "leave"
## [65] "the"          "statue"       "undraped"     "and"
## [69] "move"         "the"          "memorial"     "service"
## [73] "to"           "another"      "part"         "of"          
## [77] "the"          "museum"
```

```r
art_stories[[39]]
```

```
##   [1] "one"            "of"             "the"            "worlds"        
##   [5] "most"           "successful"     "partnerships"   "won"           
##   [9] "another"        "major"          "title"          "here"          
##  [13] "on"             "sunday"         "night"          "at"            
##  [17] "the"            "american"       "contract"       "bridge"        
##  [21] "leagues"        "summer"         "national"       "championships" 
##  [25] "the"            "life"           "master"         "pairs"         
##  [29] "which"          "has"            "a"              "#"             
##  [33] "year"           "history"        "was"            "won"           
##  [37] "by"             "a"              "wide"           "margin"        
##  [41] "by"             "robert"         "levin"          "of"            
##  [45] "riverdale"      "n"              "y"              "and"           
##  [49] "steve"          "weinstein"      "of"             "glen"          
##  [53] "ridge"          "n"              "j"              "both"          
##  [57] "have"           "won"            "the"            "event"         
##  [61] "with"           "other"          "partners"       "and"           
##  [65] "together"       "they"           "won"            "the"           
##  [69] "#"              "cavendish"      "pairs"          "in"            
##  [73] "las"            "vegas"          "these"          "were"          
##  [77] "the"            "final"          "standings"      "first"         
##  [81] "levin"          "and"            "weinstein"      "#"             
##  [85] "#"              "#"              "match"          "points"        
##  [89] "second"         "gary"           "cohler"         "of"            
##  [93] "highland"       "park"           "ill"            "and"           
##  [97] "ralph"          "katz"           "of"             "hinsdale"      
## [101] "ill"            "#"              "#"              "#"             
## [105] "third"          "robert"         "gookin"         "of"            
## [109] "falls"          "church"         "va"             "and"           
## [113] "earl"           "glickstein"     "of"             "gaithersburg"  
## [117] "md"             "#"              "#"              "#"             
## [121] "fourth"         "fred"           "stewart"        "of"            
## [125] "kingston"       "n"              "y"              "and"           
## [129] "kit"            "woolsey"        "of"             "kensington"    
## [133] "calif"          "#"              "#"              "#"             
## [137] "the"            "spingold"       "knockout"       "teams"         
## [141] "began"          "this"           "afternoon"      "with"          
## [145] "an"             "entry"          "of"             "#"             
## [149] "teams"          "the"            "top"            "seeded"        
## [153] "teams"          "by"             "captain"        "are"           
## [157] "in"             "descending"     "order"          "rose"          
## [161] "meltzer"        "of"             "los"            "gatos"         
## [165] "calif"          "nick"           "nickell"        "of"            
## [169] "manhattan"      "george"         "jacobs"         "of"            
## [173] "burr"           "ridge"          "ill"            "james"         
## [177] "cayne"          "of"             "manhattan"      "rita"          
## [181] "shugart"       "of"            "pebble"        "beach"
## [185] "calif"         "steve"         "robinson"      "of"
## [189] "arlington"     "va"            "richard"       "schwartz"
## [193] "of"            "east"          "elmhurst"      "n"
## [197] "y"             "and"           "george"        "rosenkranz"
## [201] "of"            "mexico"        "city"          "thirty"
## [205] "three"         "teams"         "are"           "entered"
## [209] "in"            "the"           "womens"        "knockout"
## [213] "teams"         "top"           "seeded"        "by"
## [217] "captain"       "are"           "kathie"        "wei"
## [221] "sender"        "of"            "nashville"     "pam"
## [225] "wittes"        "of"            "venice"        "calif"
## [229] "petra"         "hamman"        "of"            "dallas"
## [233] "jill"          "meyers"        "of"            "santa"
## [237] "monica"        "calif"         "laurie"        "vogel"
## [241] "of"            "manhattan"     "hjordis"       "eythorsdottir"
## [245] "of"            "huntsville"    "ala"           "lynn"
## [249] "baker"         "of"            "austin"        "tex"
## [253] "and"           "jan"           "martel"        "of"
## [257] "davis"         "calif"
```

Think about deleting those articles from the corpus in order to run PCA on the other ones. We change the scale of the plot in order to ignore those outliers and examine documents that are extremely simialar to each other in regard to these two principal components.

```
plot(lsi_art$x[,1:2], xlim=c(-0.02, 0.02), ylim=c(-0.01, 0.01))
```



Here we identify two closely related stories from the bundle of points that are close together on the plot. These similar pieces all seem to be reviews of works of art.

```
#identify(lsi_art$x[,1:2], n=2)
art_stories[[12]]
```

```
##   [1] "elizabeth"      "murraypaula"    "cooper"         "gallery"
##   [5] "#"              "wooster"        "street"         "at"
##   [9] "houston"        "street"         "soho"           "through"
##  [13] "april"          "#elizabeth"     "murrays"        "faith"
##  [17] "in"             "the"            "complete"       "malleability"
##  [21] "of"             "paint"          "and"            "painting"
##  [25] "continues"      "unabated"       "in"             "this"
##  [29] "small"          "impressive"     "show"           "over"
##  [33] "the"            "course"         "of"             "only"
##  [37] "five"           "canvases"       "she"            "puts"
##  [41] "her"            "art"            "through"        "its"
##  [45] "paces"          "going"          "from"           "flat"
##  [49] "to"             "shaped"         "surfaces"       "and"
##  [53] "shifting"       "her"            "semi"           "abstract"
##  [57] "narratives"     "from"           "human"          "to"
##  [61] "animal"         "to"             "architecture"   "all"
##  [65] "without"        "missing"        "a"              "beat"
##  [69] "it"             "helps"          "that"           "ms"
##  [73] "murray"         "has"            "downsized"      "things"
##  [77] "a"              "bit"            "giving"         "these"
##  [81] "works"          "a"              "compressed"     "clarity"
##  [85] "of"             "shape"          "color"          "and"
##  [89] "paint"          "handling"       "that"           "her"
##  [93] "larger"         "more"           "looming"        "efforts"
##  [97] "can"            "sometimes"      "lack"           "this"
## [101] "is"             "especially"     "the"            "case"
## [105] "with"           "arm"            "ear"            "and"
## [109] "the"            "unscrew"        "painting"       "whose"
## [113] "shaped"         "and"            "stepped"        "surfaces"
## [117] "mirror"         "each"           "other"          "but"
## [121] "whose"          "images"         "do"             "not"
## [125] "arm"            "ear"            "depicts"        "a"
## [129] "restless"       "red"            "chair"          "in"
## [133] "the"            "corner"         "of"             "a"
## [137] "red"            "yellow"         "and"            "green"
## [141] "room"           "the"            "unscrew"        "painting"
## [145] "shows"          "a"              "green"          "figure"
## [149] "sandwiched"     "between"        "two"            "mazelike"
## [153] "energy"         "fields"         "bounding"       "dog"
## [157] "the"            "largest"        "painting"       "here"
## [161] "is"             "also"           "the"            "flattest"
## [165] "and"            "one"            "of"             "the"
## [169] "few"            "conventionally" "rectangular"    "works"
## [173] "ms"             "murray"         "has"            "made"
## [177] "in"             "recent"         "years"          "framed"
## [181] "by"             "such"           "wonderful"      "details"
## [185] "as"             "a"              "fuzzy"          "magenta"
## [189] "sun"            "a"              "blue"           "tree"
## [193] "and"            "three"          "yellow"         "leaves"
## [197] "a"              "red"            "biomorphic"     "dog"
## [201] "with"           "big"            "bloblike"       "front"
## [205] "paws"           "hurtles"        "through"        "space"
## [209] "in"             "miroesque"      "slow"           "motion"
## [213] "the"            "epitome"        "of"             "canine"
## [217] "joie"           "de"             "vivre"          "the"
## [221] "scene"          "is"             "set"            "against"
## [225] "a"              "creamy"         "white"          "ground"
## [229] "whose"          "wrinkles"       "and"            "undulating"
## [233] "edges"          "suggests"       "a"              "flimsy"
## [237] "rubbery"        "surface"        "caught"         "in"
## [241] "the"            "act"            "of"             "contracting"
## [245] "like"           "a"              "balloon"        "losing"
## [249] "its"            "air"            "behind"         "this"
## [253] "luminous"       "expanse"        "which"          "confirms"
## [257] "ms"             "murrays"        "ability"        "to"
## [261] "shape"          "even"           "the"            "flattest"
## [265] "surface"        "are"            "glimmers"       "of"
## [269] "golden"         "light"          "and"            "more"
## [273] "painting"       "roberta"        "smith"
```

```
art_stories[[17]]
```

```
##   [1] "laura"          "newman"         "tenri"          "cultural"
##   [5] "institute"      "#"              "broadway"       "at"
##   [9] "prince"         "street"         "soho"           "through"
##  [13] "july"           "#laura"         "newmans"        "new"
##  [17] "paintings"      "while"          "still"          "lacking"
##  [21] "in"             "originality"    "are"            "a"
##  [25] "big"            "improvement"    "over"           "her"
##  [29] "earlier"        "landscapelike"  "abstractions"   "which"
##  [33] "often"          "seemed"         "forced"         "and"
##  [37] "heavy"          "handed"         "the"            "hints"
##  [41] "of"             "landscapes"     "remain"         "but"
##  [45] "the"            "pale"           "colors"         "and"
##  [49] "map"            "like"           "compositions"   "suggest"
##  [53] "a"              "realm"          "high"           "above"
##  [57] "the"            "earth"          "or"             "that"
##  [61] "of"             "the"            "imagination"    "a"
##  [65] "clear"          "source"         "is"             "#s"
##  [69] "new"            "image"          "painting"       "modified"
##  [73] "with"           "a"              "loose"          "childlike"
##  [77] "rendering"      "and"            "with"           "half"
##  [81] "buried"         "feminist"       "subject"        "matter"
##  [85] "that"           "seems"          "pure"           "#s"
##  [89] "one"            "can"            "imagine"        "ms"            
##  [93] "newman"         "learning"       "from"           "such"          
##  [97] "seemingly"      "unrelated"      "precedents"     "as"            
## [101] "sue"            "williams"       "and"            "dona"          
## [105] "nelson"         "in"             "phrenology"     "crudely"       
## [109] "rendered"       "circles"        "topped"         "off"           
## [113] "by"             "weird"          "blobs"          "of"            
## [117] "color"          "come"           "to"             "read"          
## [121] "as"             "the"            "heads"          "of"            
## [125] "little"         "girls"          "with"           "varying"       
## [129] "hairstyles"     "in"             "other"          "paintings"     
## [133] "these"          "heads"          "drift"          "across"        
## [137] "fields"         "of"             "scumbled"       "color"         
## [141] "or"             "in"             "pearl"          "river"         
## [145] "command"        "a"              "little"         "red"           
## [149] "rocket"         "evoking"        "nancy"          "the"           
## [153] "comic"          "strip"          "character"      "in"            
## [157] "learning"       "to"             "draw"           "fragments"     
## [161] "of"             "the"            "heads"          "outlines"      
## [165] "suggest"        "a"              "kind"           "of"            
## [169] "lexicon"        "a"              "not"            "quite"         
## [173] "formed"         "alphabet"       "nearly"         "everywhere"    
## [177] "the"            "heads"          "seem"           "to"            
## [181] "function"       "as"             "crude"          "signs"         
## [185] "of"             "female"         "psyche"         "in"            
## [189] "formation"      "meanwhile"      "the"            "paintings"     
## [193] "space"          "is"             "often"          "animated"      
## [197] "by"             "seemingly"      "finger"         "painted"       
## [201] "spirals"        "of"             "color"          "that"          
## [205] "burrow"         "across"         "the"            "surface"       
## [209] "like"           "tunneling"      "moles"          "or"            
## [213] "the"            "drawings"       "of"             "rebellious"    
## [217] "children"       "these"          "corral"         "chunks"        
## [221] "of"             "territory"      "or"             "sometimes"     
## [225] "as"             "in"             "totem"          "frame"         
## [229] "a"              "second"         "vista"          "of"            
## [233] "space"          "like"           "a"              "number"        
## [237] "of"             "women"          "painting"       "today"         
## [241] "ms"             "newman"         "seems"          "intent"        
## [245] "on"             "a"              "kind"           "of"            
## [249] "anecdotal"      "or"             "narrative"      "abstraction"   
## [253] "that"           "implies"        "meaning"        "without"       
## [257] "pinning"        "it"             "down"           "this"          
## [261] "is"             "sometimes"      "weakened"       "by"            
## [265] "a"              "formal"         "indecisiveness" "that"          
## [269] "can"            "strike"         "the"            "eye"           
## [273] "as"             "too"            "self"           "consciously"   
## [277] "unskilled"      "but"            "generally"      "ms"            
## [281] "newman"         "is"             "aiming"         "at"            
## [285] "a"               "target"         "much"           "closer"
## [289] "to"              "home"           "roberta"        "smith"
```

Documents with similar patterns of context will show up as matching the query document.

- In this vector space model, documents can inherit meaning from other documents
- If different words both score highly in the same context, the documents can still be seen by the model as similar
  - Solves synonymy - synonyms can be used in multiple documents and still point to the same context
  - Solves polysemy - “apple” farm documents and “apple” tech documents will not score in the same contexts

Naïve Bayes Text Classification

Let’s say for example:

Hypothesis = an email that you received is spam
Data (pattern of word usage in the email) = occurrence of words “sale”, “buy”, and “Viagra”
Probability email is spam given D, using Bayes Rule:

$$P(H|D) = \frac{P(H)*P(D|H)}{P(D)}$$

P(H) = prior probability of spam
P(D|H) = P(uses “sale”, uses “buy”, uses “Viagra” | spam)

Remember that joint probabilities can be written as: P(A, B) = P(A) * P(B|A), not simply P(A)*P(B).

In our case, the event an email uses “sale” is not independent from the event the email uses “buy”. But it is tedious to estimate this joint probability, especially when searching for much more indicators than just three. So for Naive Bayes, we will approximate P(D|H) using the assumption that events A and B are independent of each other.

Thus P(D|H) = P(uses “sale” | spam) * P(uses “buy” | spam) * P(uses “Viagra”| spam).

Assuming we’re given a typical hypothesis, every word can be thought of as independent of every other word. Using this simplification above, Bayes Rule is very simple to apply in text classification. Considering how inaccurate our assumptions are, the resulting probabilities are quite good at classifying documents into different categories.

More Naïve Bayes next time!