PPOL 628: Text as Data – Computational Linguistics for Social Scientists

Class 3: Text Preprocessing

Today

- Lecture: key points from readings
- Reading Discussion
- Questions about data collection
- Lab: Preprocessing.R
- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Preprocessing

- Convert input text to a numerical representation – typically a count of the number of times each unique term occurs in each document.
  - This approach is commonly referred to as a “bag of words” approach since it discards word order.
  - “the cool cat is cool” -> [“the”: 1, “cool”: 2, “cat”: 1, “is”: 1]
- Some methods of pseudo-preserving order (like n-grams).
- Contrast to methods NLP like dictionary analysis, dependency parsing, etc. that preserve order.

Tokenization

- How the input sequence of characters is split up into tokens (typically what we conceive of as words).
- Whitespace tokenization: “The brown fox” → [“The”, “brown”, “fox”]
  - How to treat multiple spaces, tabs, newlines?
  - How to treat multi-word proper nouns? E.g: “White House”
  - How to treat hyphenated terms? E.g. “new-fangled”
  - How to treat acronyms? “N.R.A” → “NRA” or [“N”, “R”, ”A”]?
- Tokenization at sentence level.
- Tokenization of compound words (e.g. German)
  - "Lebensversicherungsgesellschaftsangestellter" -> "life insurance company employee"

Punctuation

- Often punctuation/special characters are not considered to be interpretable in a bag of words setting.
  - What information does knowing the number of commas in a document give us about the document's topical content?
- Some potential exceptions: "?", "!", "☺", "2^4", "N.R.A.", etc.
  - But if we don't know what "?" was accompanying, can we interpret?
- Treat words differently if attached to punctuation?
  - "wonderful" != "wonderful."?
  - We typically separate out punctuation as its own token.
  - "Wonderful!" → ["Wonderful", "!"]
- How to handle numbers? E.g. "12,340.56" → ?

Numbers

- Whether or not a number is informative is often context dependent.
  - Knowing that a bill was referring to Section 43 of US Code could be relevant.
  - Knowing that somebody bought 8 wine glasses at the store may be irrelevant to determining what a message was about.
- Can we interpret the number out of context?
  - “Section 43” is interpretable as a token, but what about [“Section”, ”43”]?
  - May want to use n-grams approach, or something more sophisticated.
  - If documents are shorter, may be able to draw stronger conclusions.
- “Jacob ate 26 Watermelons” → [“Jacob”, “ate”, “26”, “Watermelons”]

Lowercasing

- Common step for English language texts is to lowercase all terms.
  - We probably want to treat the elephant in “Elephants rule!” and “I love elephants!” as equivalent.
  - Capitalizing first term of first word in sentence is convention.
  - “Cool elephants are cool.” → [“cool”[1]:2, “elephants”:1, “are”:1]
- Should be asking ourselves if this makes sense in other languages/character sets.
- Can be valuable as a way to distinguish proper nouns.
- Only de-capitalize first word of sentence?
  - What if sentence starts with proper noun?

Stemming

- Reduce words to their linguistic stem.
  - “party”, “partying”, and “parties” → “parti”
- Many words are of the form:
  - prefix (optional) + root + suffix (optional)
  - e.g. re + **act**, **act** + ing, re + act + ing all share common root
- Often used as a vocabulary reduction method.
- Can work well when the words you care about in your documents keep the same meaning as they change prefix/suffix.
- Can run into problems in situations like “political **party**” vs. “we love **party**ing”.

Stopword Removal

- Particularly in the case of bag of words text analysis of unigrams, some terms are not meaningful/useful/interpretable, so we sometimes remove them.
- Typically includes:
  - Function words (see next slide)
  - Domain specific stop words -- words that appear in nearly all documents (e.g. “congress” in federal legislation) or words that are not interpretable in the context of your corpus (e.g. income in a corpus of financial filings).
  - Not the same as meaningful words removed for theoretical reasons.
- Big problems if you get the stopword list wrong.
- How to handle stopwords in context of multiword expressions?

- **Determiners -- Articles:** a, an, the
- **Determiners -- Demonstratives:** that, this, those, these
- **Determiners -- Possessive pronouns:** my, your, their, our, ours, whose, his, hers, its, which
- **Determiners -- Quantifiers:** some, both, most, many, a few, a lot of, any, much, a little, enough, several, none, all
- **Conjunctions:** and, but, for, yet, neither, or, so, when, although, however, as, because, before
- **Prepositions:** in, of, between, on, with, by, at, without, through, over, across, around, into, within
- **Pronouns:** she, they, he, it, him, her, you, me, anybody, somebody, someone, anyone
- **Auxiliary verbs:** be, is, am, are, have, has, do, does, did, get, got, was, were
- **Modals:** may, might, can, could, will, would, shall, should
- **Qualifiers:** very, really, quite, somewhat, rather, too, pretty (much)
- **Question words:** how, where, what, when, why, who

https://www.thoughtco.com/function-word-grammar-1690876

https://www.ranks.nl/stopwords

a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between, both, but, by, can't, cannot, could, couldn't, did, didn't, do, does, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, he's, her, here, here's, hers, herself, him, himself, his, how, how's, i, i'd, i'll, i'm, i've, if, in, into, is, isn't, it, it's, its, itself, let's, me, more, most, mustn't, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours

https://www.ranks.nl/stopwords

a, able, about, above, abst, accordance, according, accordingly, across, act, actually, added, adj, affected, affecting, affects, after, afterwards, again, against, ah, all, almost, alone, along, already, also, although, always, am, among, amongst, an, and, announce, another, any, anybody, anyhow, anymore, anyone, anything, anyway, anyways, anywhere, apparently, approximately, are, aren, arent, arise, around, as, aside, ask, asking, at, auth, available, away, awfully, b, back, be, became, because, become, becomes, becoming, been, before, beforehand, begin, beginning, beginnings, begins, behind, being, believe, below, beside, besides, between, beyond, biol, both, brief, briefly, but, by, c, ca, came, can, cannot, can't, cause, causes, certain, certainly, co, com, come, comes, contain, containing, contains, could, couldnt, d, date, did, didn't, different, do, does, doesn't, doing, done, don't, down, downwards, due, during, e, each, ed, edu, effect, eg, eight, eighty, either, else, elsewhere, end, ending, enough, especially, et, et-al, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, except, f, far, few, ff, fifth, first, five, fix, followed, following, follows, for, former, formerly, forth, found, four, from, further, furthermore, g, gave, get, gets, getting, give, given, gives, giving, go, goes, gone, got, gotten, h, had, happens, hardly, has, hasn't, have, haven't, having, he, hed, hence, her, here, hereafter, hereby, herein, heres, hereupon, hers, herself, hes, hi, hid, him, himself, his, hither, home, how, howbeit, however, hundred, i, id, ie, if, i'll, im, immediate, immediately, importance, important, in, inc, indeed, index, information, instead, into, invention, inward, is, isn't, it, itd, it'll, its, itself, i've, j, just, k, keep, keeps, kept, kg, km, know, known, knows, l, largely, last, lately, later, latter, latterly, least, less, lest, let, lets, like, liked, likely, line, little, 'll, look, looking, looks, ltd, m, made, mainly, make, makes, many, may, maybe, me, mean, means, meantime, meanwhile, merely, mg, might, million, miss, ml, more, moreover, most, mostly, mr, mrs, much, mug, must, my, myself, n, na, name, namely, nay, nd, near, nearly, necessarily, necessary, need, needs, neither, never, nevertheless, new, next, nine, ninety, no, nobody, non, none, nonetheless, noone, nor, normally, nos, not, noted, nothing, now, nowhere, o, obtain, obtained, obviously, of, off, often, oh, ok, okay, old, omitted, on, once, one, ones, only, onto, or, ord, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, owing, own, p, page, pages, part, particular, particularly, past, per, perhaps, placed, please, plus, poorly, possible, possibly, potentially, pp, predominantly, present, previously, primarily, probably, promptly, proud, provides, put, q, que, quickly, quite, qv, r, ran, rather, rd, re, readily, really, recent, recently, ref, refs, regarding, regardless, regards, related, relatively, research, respectively, resulted, resulting, results, right, run, s, said, same, saw, say, saying, says, sec, section, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sent, seven, several, shall, she, shed, she'll, shes, should, shouldn't, show, showed, shown, showns, shows, significant, significantly, similar, similarly, since, six, slightly, so, some, somebody, somehow, someone, somethan, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specifically, specified, specify, specifying, still, stop, strongly, sub, substantially, successfully, such, sufficiently, suggest, sup, sure,t, take, taken, taking, tell, tends, th, than, thank, thanks, thanx, that, that'll, thats, that've, the, their, theirs, them, themselves, then, thence, there, thereafter, thereby, thered, therefore, therein, there'll, thereof, therere, theres, thereto, thereupon, there've, these, they, theyd, they'll, theyre, they've, think, this, those, thou, though, thoughh, thousand, throug, through, throughout, thru, thus, til, tip, to, together, too, took, toward, towards, tried, tries, truly, try, trying, ts, twice, two, u, un, under, unfortunately, unless, unlike, unlikely, until, unto, up, upon, ups, us, use, used, useful, usefully, usefulness, uses, using, usually, v, value, various, 've, very, via, viz, vol, vols, vs, w, want, wants, was, wasnt, way, we, wed, welcome, we'll, went, were, werent, we've, what, whatever, what'll, whats, when, whence, whenever, where, whereafter, whereas, whereby, wherein, wheres, whereupon, wherever, whether, which, while, whim, whither, who, whod, whoever, whole, who'll, whom, whomever, whos, whose, why, widely, willing, wish, with, within, without, wont, words, world, would, wouldnt, www, x, y, yes, yet, you, youd, you'll, your, youre, yours, yourself, yourselves, you've, z, zero

Including Multiword Expressions

- Words may have different meanings/connotations/referents in different contexts:
  - E.g. “security” has a different meaning in “social security”, “national security” and “security deposit box”
- Unigram words are often uninterpretable, whereas n-grams, phrases can be more interpretable on their own.
  - E.g. “product” vs. “product placement”, “gross national product”
- Can massively increase vocabulary size.
- Three main approaches: simple n-grams, colocation statistics, linguistic phrases.

Including Multiword Expressions

- **N-Grams**: every contiguous sequence of *n* words.
  - “my cool dog maisy” → [“my cool”, ”cool dog”, “dog maisy”]

- **Colocation Statistics**: group together words that appear together unusually often (in a statistical sense).
  - E.g. “New” + “York” tend to appear in this order very frequently so we could treat them as a single word whenever they appear together.

- **Linguistic Phrases**: sequences of words that that form semantic units, often characterized by parts of speech patterns.
  - E.g. Adjective + Noun → “heavy book”, “fast car”, “slow computer”

Infrequently Used Terms

- Decision to remove terms that appear in fewer the X number/% of documents.
- Can significantly reduce vocabulary size (Zipfs law).
- If we care about understanding document similarity, or understanding how terms relate to each other, removing infrequently used terms should theoretically not effect our analysis.
- How do we set a threshold?
- When might infrequently appearing terms matter?
- May want to collapse character classes (e.g. dollar amounts) before removing.

The Readings This Week

- Manning & Schtze, H. (1999). Foundations of Statistical Natural Language Processing. Chapter 4, Section 2.

- Denny & Spirling (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.