Andrew Paley
Intro to ML: Project Proposal

The Big, Big (Big) Idea

The Washington Post declares democracy dies in darkness. I would argue that it also doesn't do so hot in the blinding light. The inundation of constant information through the largely push-model mechanics of our current media landscape leads to an inevitable and problematic outcome: short memories and shorter attention spans, with all the attendant consequences of information overload on a democracy and a culture still dependent upon an informed citizenry.

Those citizens can't begin to keep up with even a single day's events or commentary, no matter how significant or jarring, no matter whether it involves the most notable of public figures or those in the longer tail of public life. And it's no help that so much of the information comes over-packaged, whether for the sake of infotainment cable news broadcast or the weight of ad-justifying minimum word counts. In the cacophony, much goes unnoticed -- and the information we do take in has a half-life on the order of, at best, weeks (case in point: Who among us will know the name Debbie Cox in a year, let alone remember her as the person who inexplicably moved Dodge City, Kansas' lone polling place to a surprisingly difficult-to-reach location well outside of city limits last month, or the fact that she forwarded on a subsequent email inquiry from the ACLU regarding the move with her own addition of "LOL."). It all goes down the memory hole, and with it our ability to contextualize, judge subsequent situations, and take informed action.

As we fail to keep the growing fringe of the bleeding edge of history all in our heads, we increasingly rely on the gumption of journalists-as-aggregators to push newly relevant, revived information to us the moment we need it. They're our historians in real time. So if Cox ever runs for major public office, journalists *might* dig it up. But it's not a solution that scales, especially given the ever-growing universe of data points.

Put simply, the daily news cycle generates vast amounts of information and humans are bad at scale -- the good news is that machines are generally well-equipped to help us here, if we can figure out how to use them.

So, how do I plan to use machine learning to solve democracy, news and information between now and just after Thanksgiving?

The Somewhat Smaller, Near-Term-Doable Idea

There are a bunch of app or plugin ideas I could go into -- from informed-voter apps to journalist assistants to news website plugins that provide context on the participants in a story to "give me an example of" apps for use when arguing with the extended family over the holidays -- but the underlying feature set of any app aimed at this sort of problem shares a requirement in common: the ability to model/profile public figures based on the stream of news stories about them (and providing sources for its work). These profiles could include any number of metrics, but -- given the scope of this class -- I want to start with something specific: given a topic, a public figure, and a set of news stories, can a system come up with a set of direct or paraphrased quotes by that individual pertaining to that topic?

If I'm successful quickly, the next wave could get into sentiment/position analysis on those quotes (and could even lead to some interesting clustering of individuals based on position) -- but that's definitely a stretch goal given the available time. Also, I could explore statements that aren't necessarily direct quotes (assertions of action from reputable sources, for instance). The intention is that this is a building block for future work (as per the problem statement above).

On Data

The data is going to come from https://newsapi.org, and I'll be using that API to search for and ingest stories about a specific set of predetermined topics (future versions might create new topics on the fly, but I'm trying to stay somewhat focused here) and a specific set of predetermined individuals (likely starting with members of the US Congress and Senate as a well-defined initial set, but we'll see how far it goes).

Next Steps

I need to a) come up with a list of individuals; b) come up with a list of topics; c) build a simple app to start pulling down the contents of queries against newsapi.org and sticking them in a database; d) build a baseline (likely not-incredibly-smart) quote detection system to develop a training corpus of the types of relevant snippets (leveraging some part spaCy and some part good ol' regex); and e) build a simple UI to quickly assess/declare those snippets as good or bad to improve the training corpus. Once that's all in place, I'll move on to phase two -- actually training a model to do a bunch of that work for me going forward.