Resources (/resources/)

A Beginner's Guide to word2vec AKA What's the Opposite of Canada?

By Will Critchlow (/about/people/will-critchlow/). Published January 28, 2016 (https://www.distilled.net/word2vec-examples/) in the SEO (/resources/posts/seo/) category

The idea for writing this post came from a single line in the appendix to a presentation (http://www.slideshare.net/ChristopherMoody3/word2vec-lda-and-introducing-a-new-hybrid-algorithm-lda2vec-57135994/139): "what's the opposite of Canada?". While this could be the set up for some pretty poor jokes, it's actually the entrance to a rabbit warren of fascinating geeky distractions.

It turns out that while we can typically group similar or related words together – identifying that there is some connection between "Canada" and "snow" for example – we generally have a much weaker intuition for opposites. There are obviously a relatively small set of words where we'd likely have consensus on opposites – mainly adjectives like "dark", "tall", "cold" etc. – but in general, "oppositeness" is a less well-defined concept than "similarity".

This is one of those insights that helped researchers build better machine models for understanding language. These models are important because they underpin search – and advancements in this area (https://moz.com/blog/machine-learning-for-seos) are enabling trends like the move from keywords to intents (http://www.slideshare.net/DistilledSEO/emerging-trends-in-online-search/17).

Those who've been around the industry for a while may remember Latent Dirichlet Allocation (LDA) being a topic of discussion at our first ever conference (https://moz.com/blog/lda-and-googles-rankings-well-correlated) back in 2009. LDA is one model for understanding the **topic of a page**. The approach that I want to talk about today is less about pages (or "documents" in the computer science literature) and more about **understanding words.**

Surprising and cool results from thinking of words as vectors

The particular development that I want to talk about today is a model called word2vec. If you have a mathematical or computer science background, you should head straight on over to the TensorFlow tutorial on word2vec (https://www.tensorflow.org/versions/master/tutorials/word2vec/index.html) and get stuck in. If you don't, I wanted to share some surprising and cool results that don't rely on you knowing any complex mathematics.

[Sidenote: if you haven't checked out TensorFlow (https://www.tensorflow.org/) yet, it's a really important glimpse into Google's strategy (https://stratechery.com/2015/tensorflow-and-monetizing-intellectual-property/), and worth reading more about (http://googleresearch.blogspot.co.uk/2015/11/tensorflow-googles-latest-machine_9.html)].

The easiest way to think about word2vec is that it figures out how to place words on a "chart" in such a way that their location is determined by their meaning. This means that words with similar meanings will be clustered together. This represents the intuitive part of my opening example - words with semantic relationships with them will be closer together than words without such relationships.

The surprising part, though, is that the gaps and distances on the chart turn out to have meanings as well. If you go to the place where "king" appears, and move the same distance and direction between "man" and "woman", you end up in the area of the chart where "queen" appears. This is true of all kinds of semantic relationships:

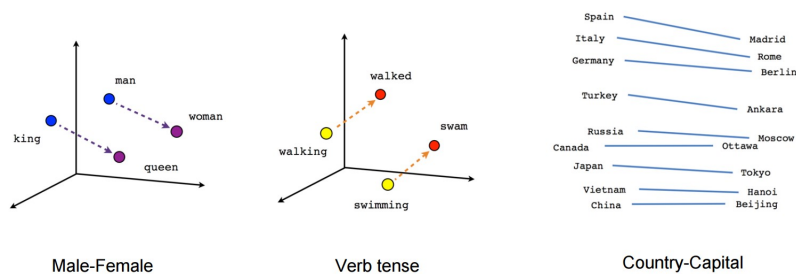Male-Female   Verb tense   Country-Capital

Image source (https://www.google.com/url?q=https://www.tensorflow.org/versions/master/tutorials/word2vec/index.html&sa=D&ust=1453985000151000&usg=AFQjCNGmFirIaXTyR0_8sdwyAy17V7LLQA)

You can think of this as:

> [king] - [man] + [woman] ~= [queen] (another way of thinking about this is that [king] - [queen] is encoding just the gendered part of [monarch])

> [walking] - [swimming] + [swam] ~= [walked] (or [swam] - [swimming] is encoding just the "past-tense-ness" of the verb)

> [madrid] - [spain] + [france] ~= [paris] (or [madrid] - [spain] ~= [paris] - [france] which is presumably roughly "capital")

You can read some more interesting examples here (http://byterot.blogspot.co.uk/2015/06/five-crazy-abstractions-my-deep-learning-word2doc-model-just-did-NLP-gensim.html) and here (https://www.quora.com/Word2vec/What-are-some-interesting-Word2Vec-results) (such as "[library] – [books] = [hall]").

Although it gets a little technical in places, this post from Stitch Fix (http://multithreaded.stitchfix.com/blog/2015/03/11/word-is-worth-a-thousand-vectors/) has some great discussion and a brilliant animated gif that explains the vector thing visually (http://multithreaded.stitchfix.com/assets/images/blog/vectors.gif).

Bringing it back to the idea that started this post ("what's the opposite of Canada?") – we might imagine "opposites" as being to do with direction of the vector – though it seems that doesn't result in entirely intuitive outcomes. This post (https://medium.com/@timsears/the-opposite-of-hitler-5eb570346020) explores this line of thought and highlights how unintuitive it is:

"If we just look for the word closest to negative hitler (-1 × hitler), we find Plantar fascia, the connective tissue that supports the arch of your foot." [source (https://medium.com/@timsears/the-opposite-of-hitler-5eb570346020)]

Some academics have discussed [warning: highly theoretical paper, PDF (http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf)] that it's no surprise that the negative of word vectors has little intuitive meaning because humans aren't really equipped with a general concept of "opposite".

So what else can we do with these vectors?

Well, it turns out that given the right corpus of text to learn from, you can do word-by-word translation [PDF (http://arxiv.org/pdf/1309.4168.pdf)]. In other words, you can't expect to translate a document – because the model has no concept of sentences or context, but you can ask "what is the Spanish word for 'one'?". The amazing thing is that in the vector space, this amounts to a rotation – in other words the relationship between words is roughly preserved across language:
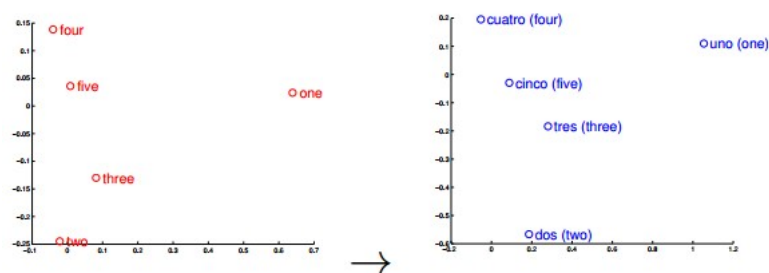
Applications

I tend to learn about this stuff in large part to understand the cutting edge of what computer language processing is capable of – I find it gives me a better understanding of how Google might be working, and where the limits might be. This is a bigger piece of the puzzle for me than actually finding real-world applications where I can use this stuff.

It's interesting to note for example that a Google spokesperson said, of RankBrain (http://www.thesempost.com/rankbrain-everything-we-know-about-googles-ai-algorithm/):

"It's related to word2vec in that it uses 'embeddings' - looking at phrases in high-dimensional space to learn how they're related to one another."

Having said that, there are applications for this kind of technology in building compelling websites at scale. Two examples I've come across are:

Dice building scalable skills pages

Dice uses it (http://www.slideshare.net/lucidworks/implementing-conceptual-search-in-solr-using-lsa-and-word2vec-presented-by-simon-hughes-dicecom) to power skills pages. They use a word vector model to understand that, for example, analytics (https://www.dice.com/skills/Analytics.html) should be classified as similar to business intelligence. This allows them to group together job adverts even if they don't use the exact same language to describe the role.

Stitch Fix classifying items of clothing

The afore-mentioned post from Stitch Fix (http://multithreaded.stitchfix.com/blog/2015/03/11/word-is-worth-a-thousand-vectors/) describes how they use this technology to build a recommendation engine that can, for example, find a top like this one:

That is suitable for someone who is pregnant:

Concerns

There are a variety of theoretical challenges and constraints that come with this kind of approach - the biggest being that context is a huge part of language, and any "bag of words" model will be limited in its usefulness. I'm inclined to say that the biggest leaps forward from here are more likely to be document models than word models, but this stuff is still fascinating.

The broader concern is that any kind of machine learning is only as good as its source information (the "corpus" of text in this case). Specifically, when we're talking about understanding language, we have to realise that any biases embedded in the corpus will remain in the model. If the training set has sexist language, for example, the machine will make sexist associations. It's ridiculously easy to find spurious correlations (http://tylervigen.com/discover) in large datasets - and we know that existing humans and human language have a ton of biases. Without a lot of care, we could easily end up in a situation where our supposedly rational machines encode these biases. Techcrunch had an interesting article on the dilemma (http://techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasy-pair/).

Recommended Reading

If you've enjoyed this high-level overview, and have more of an academic / mathematical mind-set, you might like to dive into some of the more theoretical areas:

› Quora (https://www.quora.com/How-does-word2vec-work) thread linking to a variety of interesting articles

› The original announcement from Google back in 2013 (http://google-opensource.blogspot.co.uk/2013/08/learning-meaning-behind-words.html) when they first open-sourced the code

› TensorFlow introduction to the word2vec algorithms and tutorial (https://www.tensorflow.org/versions/master/tutorials/word2vec/index.html)

› Generalising word2vec into a document classifier (http://eng.kifi.com/from-word2vec-to-doc2vec-an-approach-driven-by-chinese-restaurant-process/)

Get blog posts via email

Email address...

Subscribe

About the author

Will Critchlow (/about/people/will-critchlow/)
Will founded Distilled with Duncan in 2005. Since then, he has consulted with some of the world's largest organisations and most famous websites, spoken at most major industry events and regularly appeared in local and national press. For the... read more (/about/people/will-critchlow/)

(/)
Consulting & Services (/services/)

Conferences (/events/)

Resources (/resources/)

Privacy Policy (/privacy-policy/)

Contact (/contact/)

Get the best of Distilled in your inbox

Email address   Subscribe

Follow us

f (http://www.facebook.com/distilled)   (http://twitter.com/distilled)   (/feed/)