Distributional similarity and how to evaluate your models

Emiel van Miltenburg

July 1, 2015

Recap relatedness & similarity

- Relatedness: association between items.
- Similarity: how much two items are alike.

Evaluation

- Test relatedness using MEN data set.
- Test similarity using SimLex data set.

#1 Rule

Everything we do for English, we should also do for Dutch.

So yes...



What I've made

- Two corpora (Dutch government, Dutch Wikipedia)
- Dutch Word2Vec models (using above + NLCOW*)
- Dutch test bench for similarity & relatedness
- Demo to show the capabilities of those models

^{*} A parser for COW corpora is available at https://github.com/evanmiltenburg/cowparser

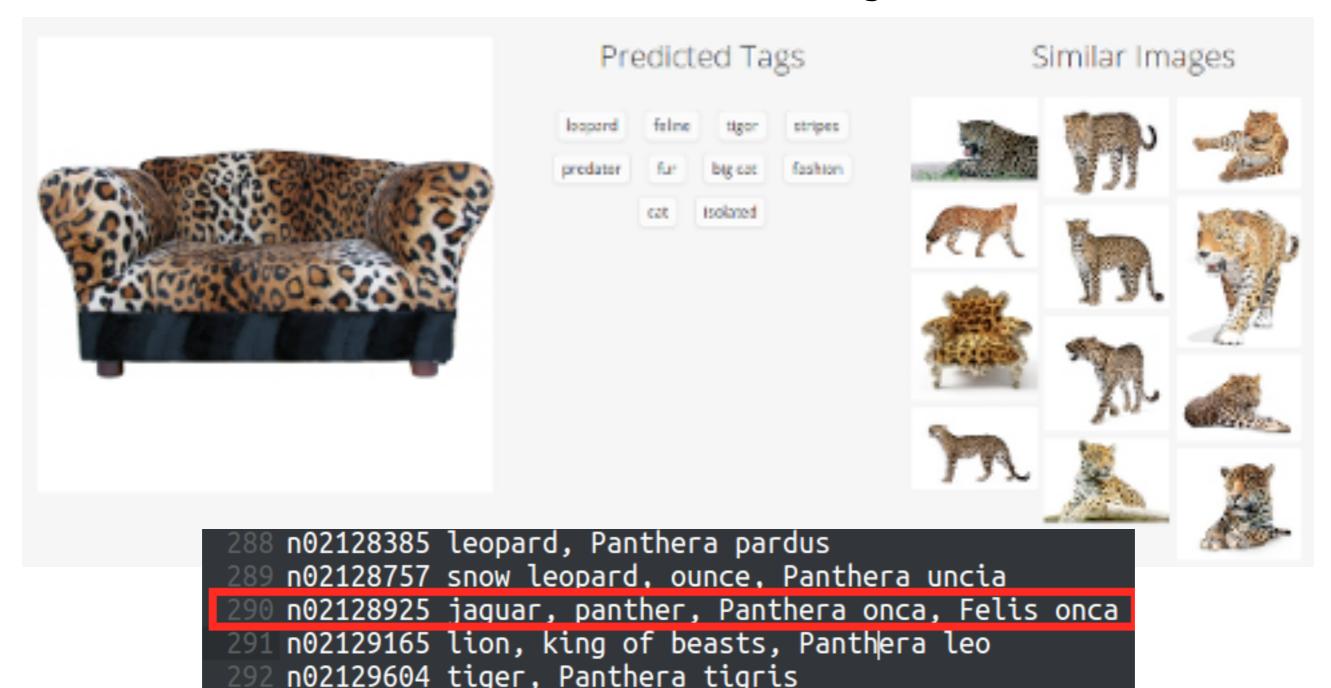
But there's more...

Remember last week's email?



"Why are you fooling around with those strange animal pictures?"

Here's why:



http://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html

#2 Rule

Models should be doing more than simple pattern matching: they should bring us closer to *understanding*

What are these animals?



Humans can...

- Recognize the different animals that make up a particular hybrid
- Recognize that this is not a case of occlusion

•