



fountain

Natural Language Processing:
Text Classification in Scala

<https://github.com/apatzer/basic-open-nlp.git>

Aaron Patzer, CEO (aaron@fountain.com)

Jean Sini, CTO (jean@fountain.com)



Why use NLP?

- Categorize spending (Mint.com)
- Sentiment {Positive, Neutral, Negative}
- Spam detection
- Question Answering (auto-chat)
- Document summarization
- Inferring implicit skills (Fountain.com)



Understanding Language

Break sentences / documents into manageable parts

Original	My Moen faucets are all leaking
Lemmetization (smart stemming)	My, Moen, faucet, be, all, leak
Parts of Speech	PRP, NNP, NN, VBP, DET, VBG
Entities	O, Organization, O, O, O, O
Dependency Graph	



Turn Documents into Feature Vectors

Words	Word Endings	Word Pairs	Skills Found	Entities
S-my	SE-cet	SWP-my-moen	SK-faucets	ORG-moen
S-moen		SWP-moen-faucet		
S-faucet		SWP-faucet-leak		
S-leak				

Training data => Vectors become a Matrix

Convex optimization balances conflicting data



	Plumbing	Gardening	...	Home Repair	Software
S-faucet	1.75	0.65		0.43	-0.45
SWP-faucet-leak	2.35	0.12		0.21	0
ORG-moen	0.45	2E-02		0.70	0
S-leak	1.01	0.11		0.45	0.92



Score an unknown using the Matrix

“My Moen faucets are all leaking”

- Score each category, normalize, then sort:
 $P(\text{Plumbing}) = e^{1.75} + e^{2.35} + e^{0.45} + e^{1.01}$
 $P(\text{Software}) = e^{-0.45} + e^{-1.97} + e^0 + e^{0.92}$
- Result is something like:
 $\{(\text{Plumbing}, 0.86), (\text{Gardening}, 0.12), (\text{Home Repair}, 0.01) \dots\}$
- Probabilities will add to ~ 1.0