A Dictionary of Wisdom and Wit: Learning to Extract Quotable Phrases

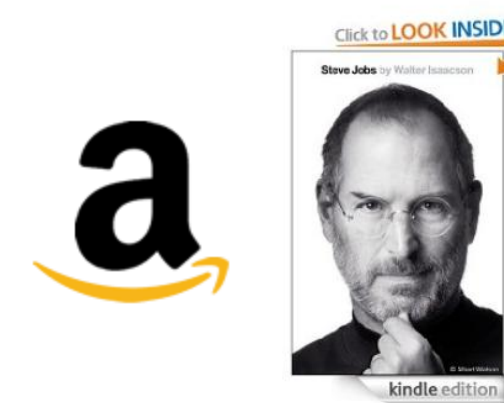**Michael Bendersky**   *University of Massachusetts Amherst*   David A. Smith

Quotable Phrase

– A meaningful, memorable, and succinct statement that can be quoted without its original context
– But, *not necessarily an already famous quote*

Motivation

– Automatically extract quotable phrases for book promotion and advertisement

Popular Highlights (What's this?)

"'Pretend to be completely in control and people will assume that you are.'"
Highlighted by 3886 Kindle users

"He emphasized that you should never start a company with the goal of getting rich.
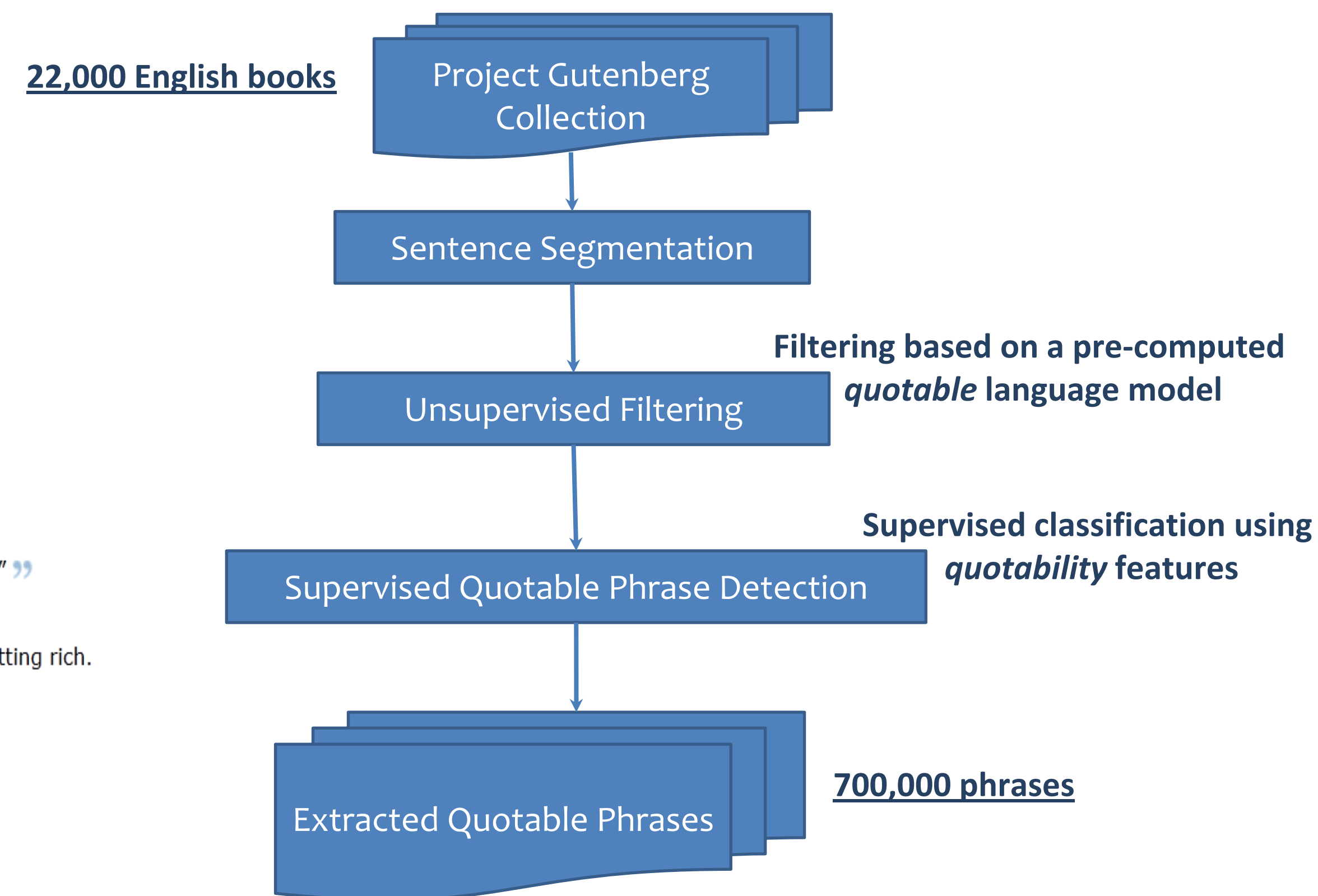Highlighted by 2933 Kindle users

Our System

• "You know that the love of money is the root of all evil."
• "I wish to live a life worth living, wherein is pleasure and pain, the lack of sameness, and the joy of things new."

– Re-discover forgotten authors and understand what makes a phrase quotable

> "There is life in a poet so long as he is quoted."
>
> *Alfred Comyn Lyall,*
> *"Studies in Literature and History"*

Visit **noisypearls.com** for more details about the project

Quotable Phrase Extraction

22,000 English books → Project Gutenberg Collection

↓

Sentence Segmentation

↓

Filtering based on a pre-computed *quotable* language model

Unsupervised Filtering

↓

Supervised classification using *quotability* features

Supervised Quotable Phrase Detection

↓

Extracted Quotable Phrases — **700,000 phrases**

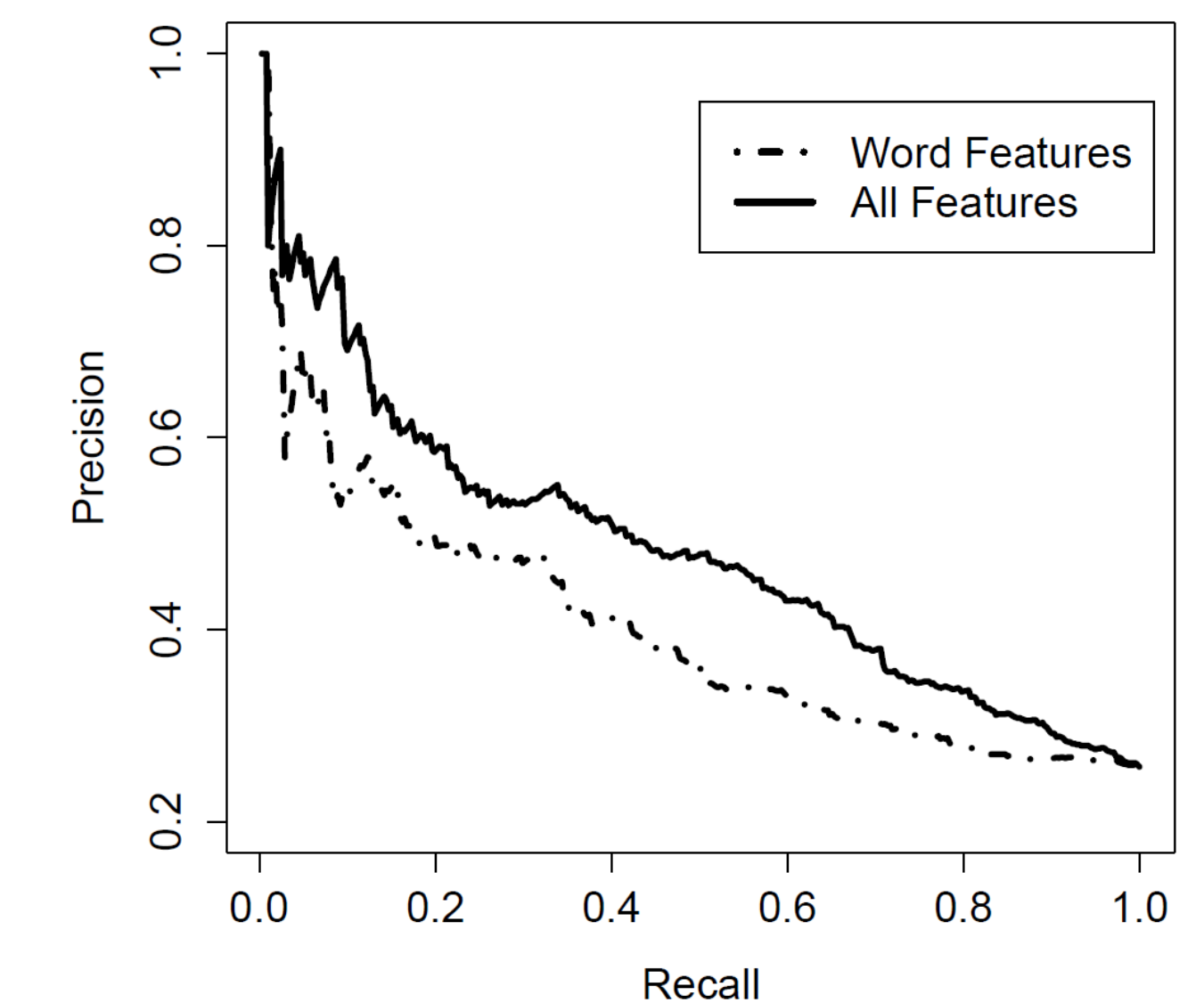## Quotability Feature Types

– **Words** – quote length, capitalization patterns, presence of universal quantifiers and abstract concepts in the quote
– **Punctuation** – presence of different punctuation types in the quote
– **Parts of Speech** – parts of speech patterns in the quote

Top Quotability Features

Feature Description	Score
# abstract concepts	+ 91.64
# universal quantifiers	+ 61.67
contains a verb in Past Participle	- 60.34
# POS sequences [VB IN PRP]	+ 39.71
# POS sequences [PRP MD VB]	- 38.78
# adjectives	+ 37.71
# POS sequences [DT NN VBD]	-36.88
# verbs	+35.22

Quotable Phrase Detection – PR Curve

reddit Study

Distribution of the *reddit* upvote (⇑) scores

Popular	$\Uparrow \geq 10$	12
Upvoted	$1 \leq \Uparrow \leq 10$	34
No upvotes	$\Uparrow \leq 0$	14
$\mathbf{p(\Uparrow > 0)} =$		**.77**

Corpus Analysis

Quotabilty Index $QI(x) = \dfrac{\# \text{ quotable phrases}(x)}{\# \text{ total sentences}(x)}$

Authors/Books with the highest Quotability Index

1	Henry Drummond	1	"Friendship" (Hugh Black)
2	Ella Wheeler Wilcox	2	"The Dhammapada" (Translated by F. Max Muller)
3	S. D. Gordon	3	"The Philosophy of Despair" (David Starr Jordan)
4	Andrew Murray	4	"Unity of Good" (Mary Baker Eddy)
5	Ralph Waldo Emerson	5	"Laments" (Jan Kochanowski)
6	Orison Swett Marden	6	"Joy and Power" (Henry van Dyke)
7	Mary Baker Eddy	7	"Polyeucte" (Pierre Corneille)
8	'Abdu'l-Bahá	8	"The Forgotten Threshold" (Arthur Middleton)
9	John Hartley	9	"The Silence" (David V. Bush)
10	Rabindranath Tagore	10	"Levels of Living" (Henry Frederick Cope)