Idea Density – A Potentially Informative Characteristic of Retrieved Documents

Michael A. Covington
Institute for Artificial Intelligence

The University of Georgia

Institute for Artificial Intelligence
The University of Georgia
1785

caspr
COMPUTER ANALYSIS OF SPEECH FOR PSYCHOLOGICAL RESEARCH

Introduction

We usually judge information retrieval by whether it finds documents on the right subject.

But the type of document is also important.

Introduction

This preliminary study indicates that idea density can help tell you whether a document is written for popular or specialized audiences.

What is idea density?

Idea density =

Number of propositions

÷ Number of words

What is idea density?

Propositions =

information =

whatever can be true or false.

What is idea density?

Example:

The old gray mare has a big nose.

Propositions:

1. **Mare is old.**
2. **Mare is gray.**
3. **Mare has nose.**
4. **Nose is big.**

4 propositions ÷ 8 words = 0.500 idea density

What is idea density?

Low idea density =

“short, choppy sentences” =

relatively little information per sentence.

The mare is old, the mare is gray…

(Idea density = 0.250, very low)

What is idea density?

High idea density =

dense packing of information =

complex interrelationships expressed.

The gray mare is very slightly older than…

(Idea density = 0.625, very high)

What is idea density?

Idea density is used extensively in studies of reading comprehension and memory (Kintsch, 1974, 1998).

Low idea density in speech or writing can indicate mental disorders, including Alzheimer's disease (Snowdon et al. 1996; Covington et al. 2007).

What is idea density?

Idea density, by now, a traditional psycholinguistic measurement.

A case can be made for bringing it into line with modern semantic theory…

…but usual practice (including ours) is to replicate Kintsch's traditional rating method (and Turner & Greene's examples).

Methodology

In this study, 14 documents were retrieved, all on the subject of U.S. monetary policy:

10 answers to Google query "predict U.S. inflation rate"

+

4 speeches or reports by Fed chairmen Bernanke and Greenspan

Methodology

Prior to analysis, the 14 texts were classified into 4 types:

Popular (news media)

Introductory (Wikipedia, Investopedia)

Scholarly (refereed journals)

Technical (policymaker-to-policymaker)

Methodology

Idea density of all documents was measured using CPIDR software developed at UGA (Brown et al. 2008).

CPIDR uses part-of-speech tagging and pattern matching to achieve high accuracy without full parsing.

It was calibrated against Turner and Greene's idea density benchmarks.

Methodology

CPIDR rates idea density using a 2-step process:

(1) Part-of-speech tagging

(2) Readjustment rules to correct the handling of certain configurations of words

Verbs, prepositions, adjectives, adverbs, conjunctions are usually propositions;

nouns, pronouns, and determiners are not.

Methodology

Example of low idea density

An increase in the factory workweek made the biggest contribution…

- Bloomberg News

(“Nouny” style = low idea density)

Methodology

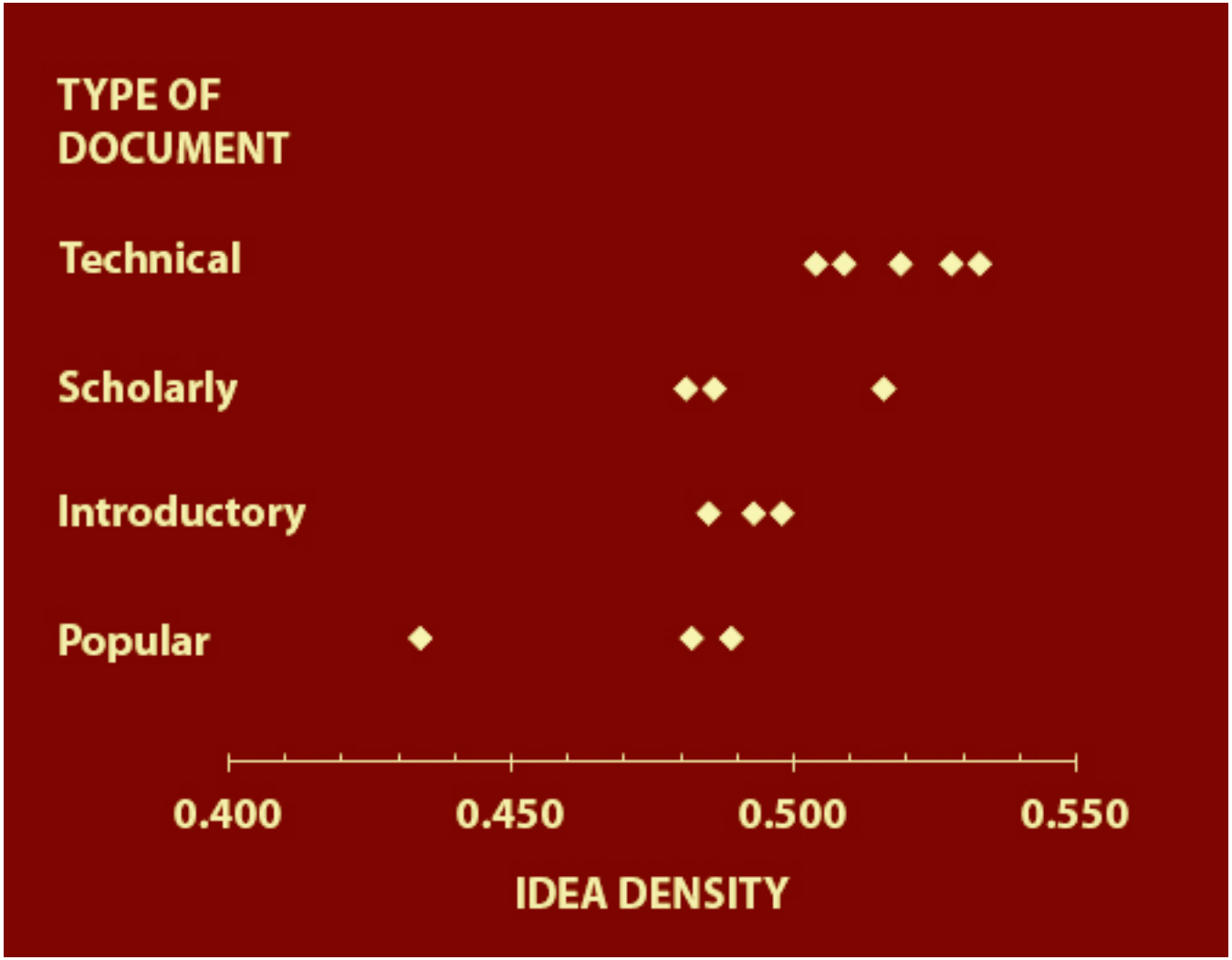
Example of high idea density

…they perceive less risk than they do for objectively comparable investments…

- Alan Greenspan

(Lots of description, comparisons, and qualifiers)

Results



Clearly, idea density discriminates document types.

Results

So far so good.

But are we just measuring "reading level"?

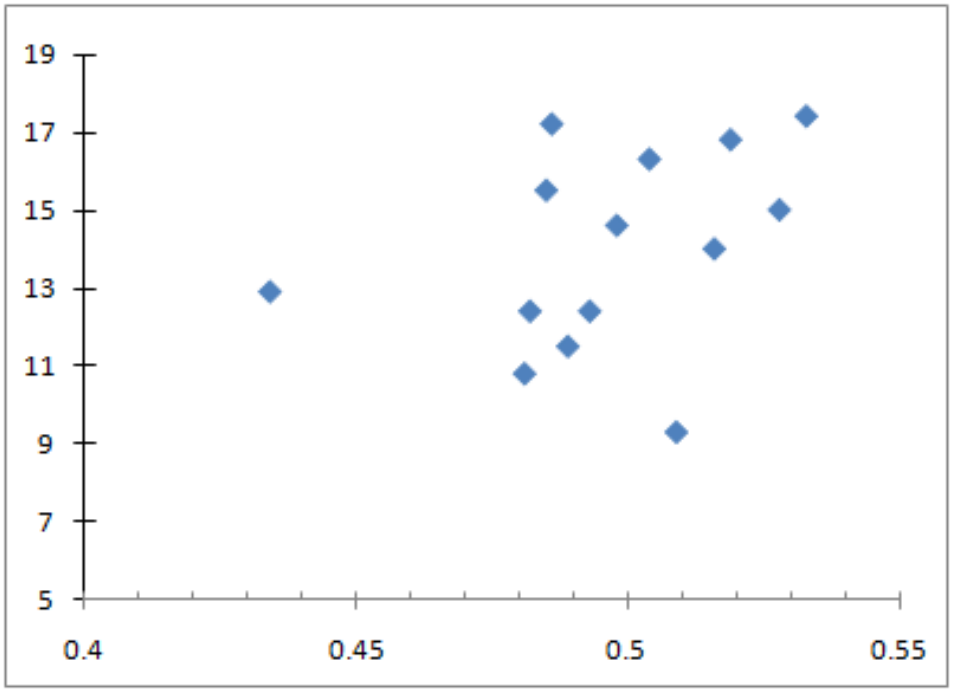
Or are we really onto something new?

Results

Idea density (CASPR) does not correlate with Flesch-Kincaid reading level (Microsoft *Word*)…

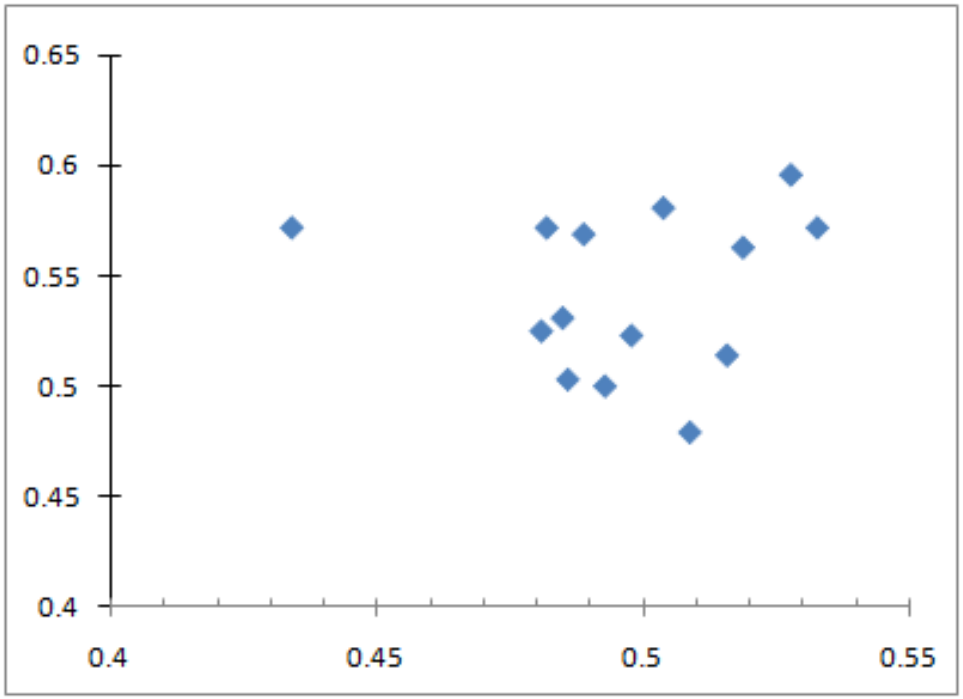
***r* = 0.356**

***P* = 0.21**



Results

…nor with vocabulary size (as indicated by average type-token ratio of a 300-word moving window)…



***r* = 0.053**

***P* = 0.85**

Results

Conclusion:

Idea density is a new, different, and useful measurement of whether a text is popular, introductory, or technical.

Results

To do next:

Replicate this study with larger sets of texts and more sophisticated evaluation criteria.

?