Novelty Detection in Text Corpora
An Illustration from Case Study 1

Hillary Clinton Controversies, Spring-Summer 2015

Text Analytics (PREDICT 453)
Northwestern University
School of Professional Studies
Master of Science in Predictive Analytics
Summer, 2015

NORTHWESTERN UNIVERSITY
SCHOOL OF PROFESSIONAL STUDIES

Alianna J. Maren, Ph.D.
alianna.maren@northwestern.edu
alianna@aliannajmaren.com
www.aliannajmaren.com

Novelty Detection:

Hillary Clinton and Alleged Classified Emails

How Do We Discern Novelty in the New Document?

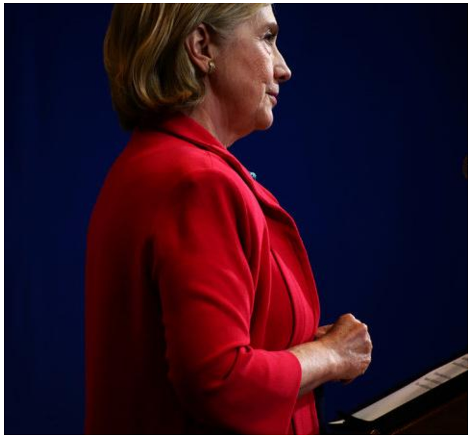
Prototype for Precursor Document Set

"Hillary Clinton emails: Did she do anything wrong or not?"

Jeremy Diamond and Elise Labott, CNN
Weds., March 11, 2015
http://www.cnn.com/2015/03/06/politics/hillary-clinton-emails-was-there-wrongdoing/

New Document:

"Hillary Clinton emails said to contain classified data"



Michael S. Schmidt and Matt Apuzzo
The New York Times
July 24, 2015
http://www.nytimes.com/2015/07/25/us/politics/hillary-clinton-email-classified-information-inspector-general-intelligence-community.html?_r=0

Step 1: Find the Most-Similar Set of Documents

New document

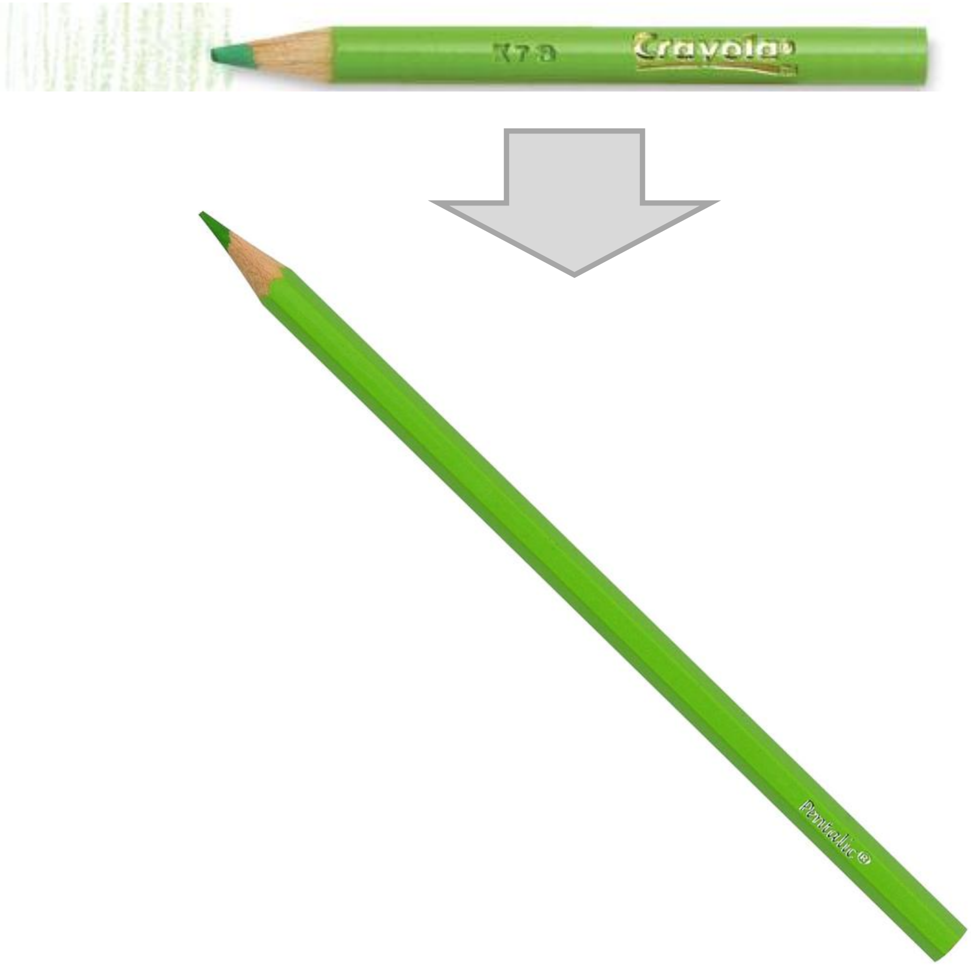
Compare with SETS of similar documents

Each Set of Similar Documents Has a Prototype Vector

Small Set of
Prototype Vectors

Documents Grouped According
to Best-Matching Prototype

Start by Finding the Best-Matching Prototype Vector



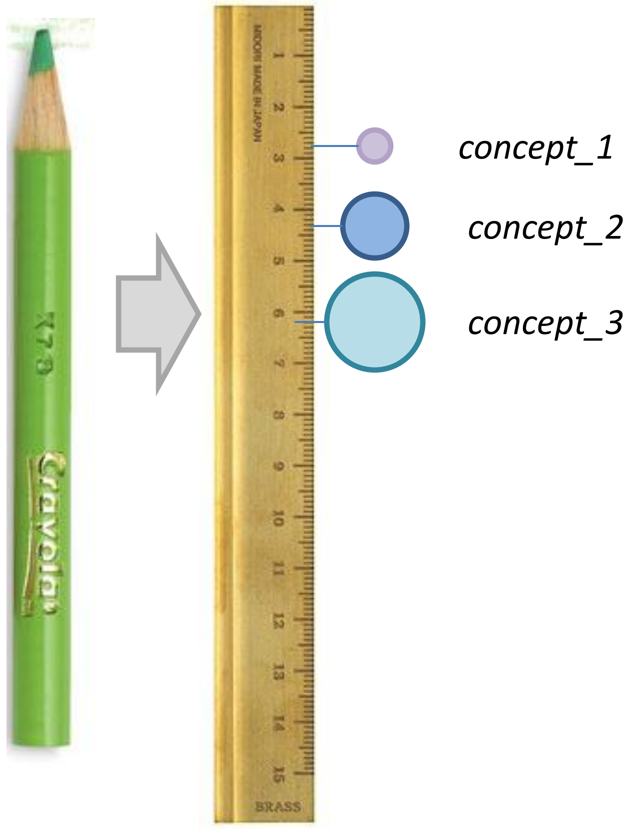
Then: ***Find Out What Is New and Different***

Document Matching: Using Vector-Matching to Find Document Similarity

The Big Question: Match Using Terms or Concepts?

- Hundreds of terms per document
- Dozens of concepts per document
- ***Concepts are more concise representation***

Vector representation of document content



Document matching uses vector similarity

- **Documents are similar if components are:**
  - Similar in nature
  - Similar in relative strength
- **HOWEVER**, vector matching algorithms require ***matching <u>vector element fields</u> => a TOUGH CONSTRAINT!***

It is easier to match documents using concepts than terms

- ***Concepts condense relevant terms*** into more compact and precise units
- ***Concepts are more general***, and many terms contribute to each concept
- ***Concept-matching has less error*** than term-matching when determining document similarity

Example:

Document Matching and Novelty Detection

Starting Point:
An Exemplar Document:

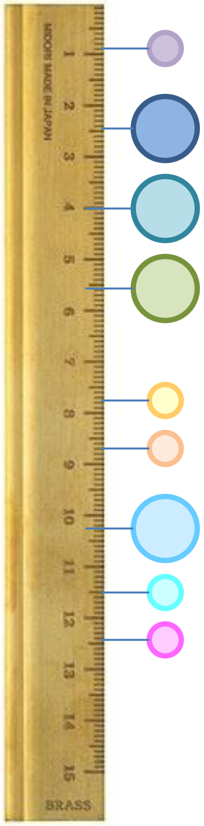
"Hillary Clinton emails: Did she do anything wrong or not?"

Jeremy Diamond and Elise Labott, CNN
Weds., March 11, 2015
http://www.cnn.com/2015/03/06/politics/hillary-clinton-emails-was-there-wrongdoing/

Document Concepts: *Vector representation of document content*

***Concepts** found in this document*

(illustrative subset)



hillary_clinton

hillary_clinton_sec_state

hillary_clinton_sec_state_controversy

hillary_clinton_sec_state_controversy_email

dept_state

dept_state_policies

email_private_server

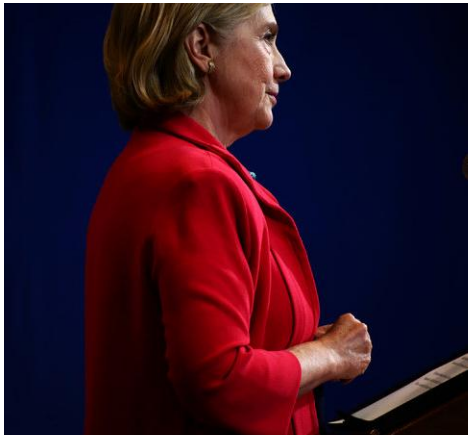
email_govt_server

sensitive_but_unclassified

Novelty Detection: Discerning "Significant Newness"

New Document:

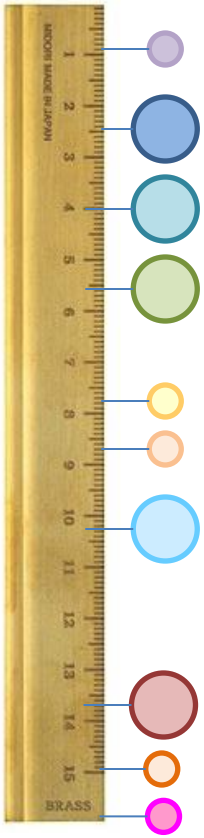
"Hillary Clinton emails said to contain classified data"



Michael S. Schmidt and Matt Apuzzo
The New York Times
July 24, 2015
http://www.nytimes.com/2015/07/25/us/politics/hillary-clinton-email-classified-information-inspector-general-intelligence-community.html?_r=0

Document Concepts: *Vector representation of document content*

***Concepts** found in this document*
(illustrative subset)



hillary_clinton

hillary_clinton_sec_state

hillary_clinton_sec_state_controversy

hillary_clinton_sec_state_controversy_email

dept_state
dept_state_policies

email_private_server

email_govt_server
sensitive_but_unclassified

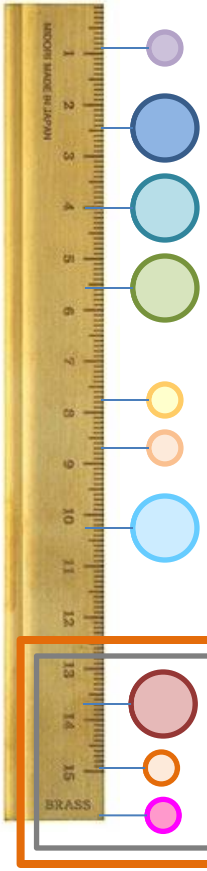
classified

inspectors_general
intelligence_agencies

Novelty Detection:

Compare New Document Concepts Against Prototype

New Document Concepts



hillary_clinton

hillary_clinton_sec_state

hillary_clinton_sec_state_controversy

hillary_clinton_sec_state_controversy_en

dept_state

dept_state_policies

email_private_server

email_govt_server

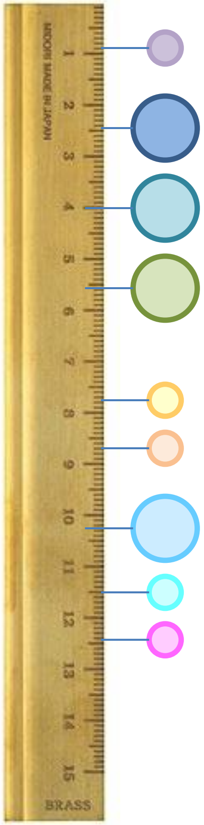
sensitive_but_unclassified

classified

inspectors_general

intelligence_agencies

Closest Match: Prototype Document Concepts



hillary_clinton

hillary_clinton_sec_state

hillary_clinton_sec_state_controversy

hillary_clinton_sec_state_controversy_email

dept_state

dept_state_policies

email_private_server

email_govt_server

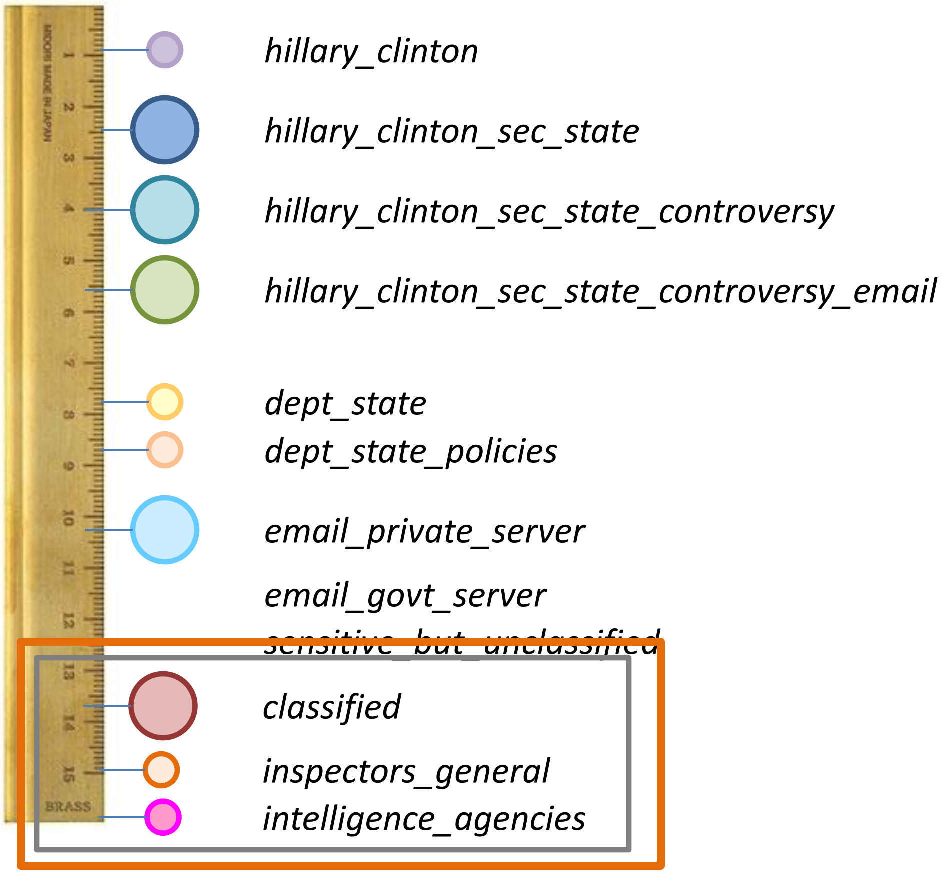
sensitive_but_unclassified

New concepts *(or even unassigned terms)* versus prototype

Summary

Novelty Detection Happens When We Find Difference Against Known Prototypes

New Document Concepts



hillary_clinton

hillary_clinton_sec_state

hillary_clinton_sec_state_controversy

hillary_clinton_sec_state_controversy_email

dept_state

dept_state_policies

email_private_server

email_govt_server

sensitive_but_unclassified

classified

inspectors_general

intelligence_agencies

Summary of Steps

- **Find closest prototype match**
- **Find key differences**
- **Test for strength**
    - Numbers of similar new documents
    - Strengths of new terms
- **Threshold to flag novelty**

In This Particular Case

- **Breaking news, multiple channels**
- **Consistent use of new terms: "classified," "inspectors general," "intelligence agencies"**
- **Significant difference from prototype: "sensitive but unclassified"**

Result: Novel Terms Detected

The Crucial Question

We detected novelty because we matched a document against its best-matching prototype and found differences

We matched using concepts, rather than terms, because it gave us a more accurate prototype match

So the BIG QUESTION is:
How do we go from terms to concepts?

The ANSWER:
This is the <u>toughest task</u> in text analytics.
Look for the SEQUEL – coming soon!