

NLP*

from Strings to Things

* [natural language processing](#)

But it has limitations



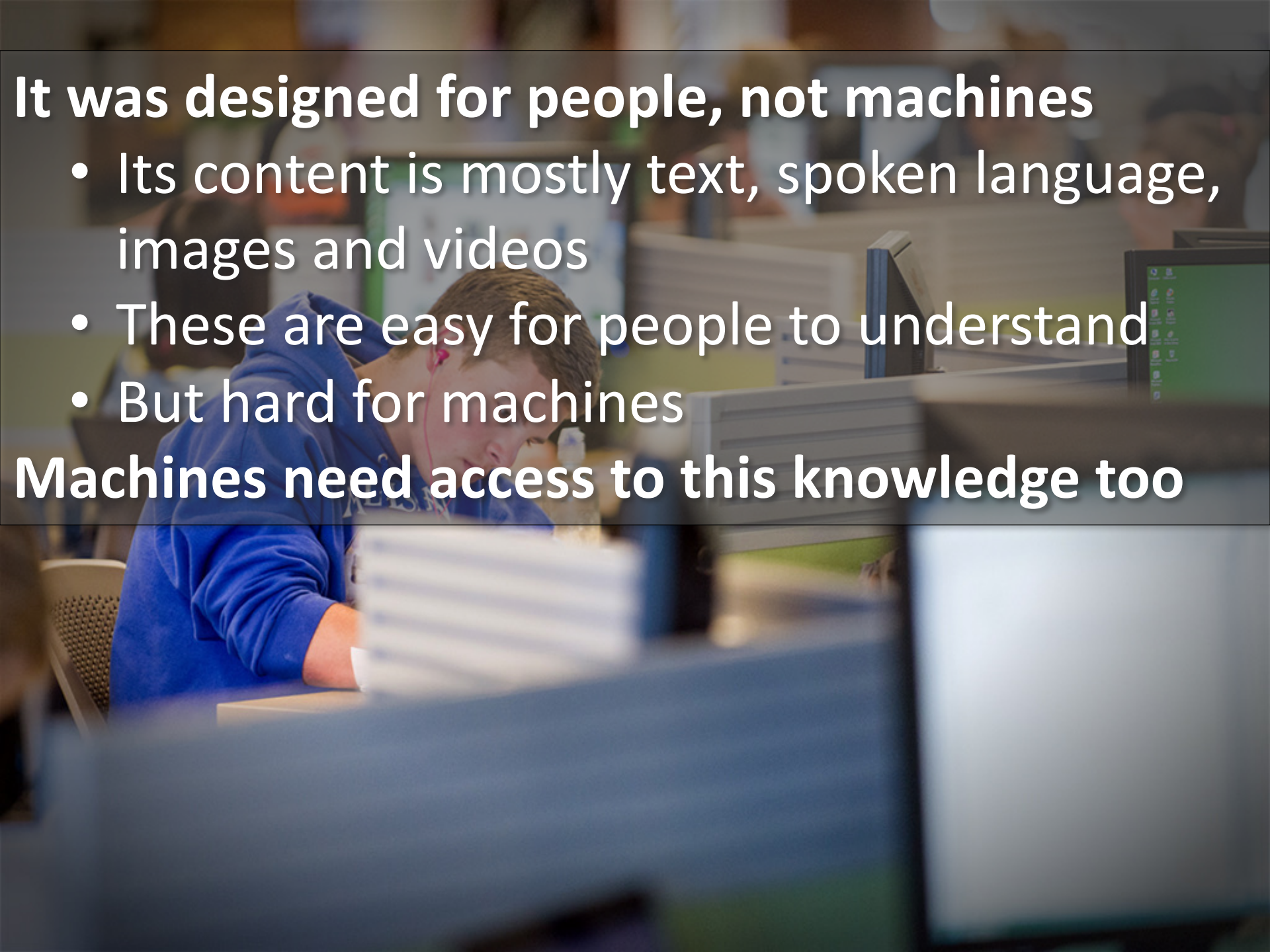
It was designed for people, not machines



It was designed for people, not machines

- Its content is mostly text, spoken language, images and videos
- These are easy for people to understand
- But hard for machines

Machines need access to this knowledge too



Access is primarily via information retrieval



Vannevar Bush envisioned a hypertext/IR system in 1945

Access is primarily via information retrieval

- Key-word queries → ranked document list
- We still need to read the documents or watch the videos
- We often want an answer to a question

And so do our machines and apps

Vannevar Bush envisioned a hypertext/IR system in 1945

We need to add knowledge graphs





We need to add knowledge graphs

- High quality semi-structured information about entities, events and relations
- Represented & accessed via standard APIs
- Easily integrated, fused and reasoned with



State of the Art?

Google is a good example, but Microsoft, IBM, Apple and Facebook all have similar capabilities

- 2010 Google acquired MediaWeb and its **Freebase** KB
- 2014: Freebase: 1.2B facts about 43M entities
- 2015+: Google knowledge graph, updated by text IE

DBpedia open source RDF KB is another

- 800M facts about 4.6M subjects from English **Wikipedia**, data also available in 21 other languages
- Helps integrate 90B facts from 1000 RDF datasets in the linked data cloud

Wikidata Knowledge Graph

- Large knowledge graph with 1B statements about ~72M items
- Fine-grained ontology: ~2M types; ~5K properties
- Multilingual, strings tagged with language id
- Links to entity's **Wikimedia pages**
- Entities have a canonical **name** and **aliases** in multiple languages and multiple claims
- UMBC=Q64780099, with type University, 569 statements
- Editable by humans and bots
- Can query with SPARQL query language

University of Maryland, Baltimore County (Q735049)

public university in Maryland
UMBC

- In more languages
Configure

Language	Label	Description	Also known as
English	University of Maryland, Baltimore County	public university in Maryland	UMBC
Spanish	No label defined	No description defined	
Traditional Chinese	馬里蘭大學巴爾的摩縣分校	No description defined	
Chinese	马里兰大学巴尔的摩县分校	No description defined	馬里蘭大學巴爾的摩縣分校


All entered languages

Statements

- instance of university edit
• 1 reference
- public educational institution of the United States edit
• 1 reference
- research university edit
• 1 reference

+ add value

logo image edit



University of Maryland, Baltimore County
logo.svg
512 × 118; 7 KB

[Q64780099](#)

Ask: When was Tom Sawyer written?

The screenshot shows a Google search for "when was tom sawyer written". The search bar at the top contains the query. Below the search bar, the results show "About 5,780,000 results (1.06 seconds)". A red box highlights the answer: "The Adventures of Tom Sawyer / Date written" followed by "1875". Below this, there is a section for "People also search for" with three items: "Adventures of Huckleb... 1883", "Treasure Island 1881", and "Robinson Crusoe 1719". To the right, a "Google Knowledge Graph data" panel is highlighted with a red border. It displays the book cover for "The Adventures of Tom Sawyer" by Mark Twain, a "Book preview" section with a "READ NOW" button, and a section stating "82% liked this book" with thumbs up and down icons. Below that, a paragraph of text describes the book: "The Adventures of Tom Sawyer is an 1876 novel by Mark Twain about a boy growing up along the Mississippi River. It is set in the 1840s in the town of St. Petersburg, which is based on Hannibal, Missouri, where Twain lived as a boy. [Wikipedia](#)". At the bottom of the panel, it lists "Originally published: June 1876", "Author: [Mark Twain](#)", "Original language: English", and "Genres: Novel, Children's literature, Satire, Adventure fiction, Folklore, Bildungsroman, Discontinuous novel".

when was tom sawyer written

when was tom sawyer written

All Images News Videos Maps More Tools

About 5,780,000 results (1.06 seconds)

The Adventures of Tom Sawyer / Date written

1875

answer

People also search for

Adventures of Huckleb... 1883

Treasure Island 1881

Robinson Crusoe 1719

Feedback

People also ask

What year is Tom Sawyer set in?

Why did Mark Twain wrote Tom Sawyer?

Who was Tom Sawyer based on?

How old is Tom Sawyer in the book?

Feedback

Google Knowledge Graph data

The Adventures of Tom Sawyer

Novel by Mark Twain

Book preview

Full book available

READ NOW

82% liked this book

Google users

The Adventures of Tom Sawyer is an 1876 novel by Mark Twain about a boy growing up along the Mississippi River. It is set in the 1840s in the town of St. Petersburg, which is based on Hannibal, Missouri, where Twain lived as a boy. [Wikipedia](#)

Originally published: June 1876

Author: [Mark Twain](#)

Original language: English

Genres: Novel, Children's literature, Satire, Adventure fiction, Folklore, Bildungsroman, Discontinuous novel



BROWSE ▾

Find a recipe

Ingredient Search



Create a profile

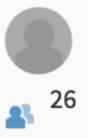


Home > Recipes > Desserts > Pies > Fruit Pies

Apple Pie by Grandma Ople




9K made it | 6969 reviews



Recipe by: MOSHASMAMA

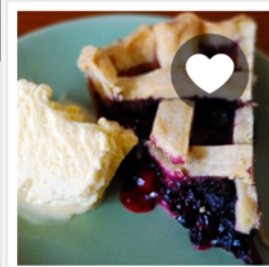
"This was my grandmother's apple pie recipe. I have never seen another one quite like it. It will always be my favorite and has won me several first place prizes in local competitions. I hope it becomes one of your favorites as well!"



>  **Grandma Ople's Apple Pie**
 ★★★★★ 1930

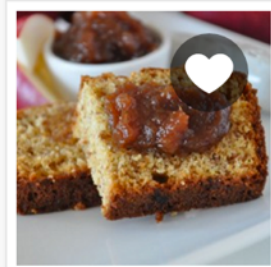
Related

Recipes Videos Categories Articles



Blueberry Pie
★★★★★ 1K

Recipe by ASHESP
3 hearts 1 flag



All-Day Apple Butter
★★★★★ 883

Recipe by Terri

Featured in Allrecipes Magazine



Save



I Made It



Rate it



Share



Print

Ingredients

1 h 30 m ⌚ 8 servings 🍽 512 cals 📊

+ 1 recipe pastry for a 9 inch double crust pie

+ 1/2 cup white sugar

+ 1/2 cup unsalted butter

Domino Pure Cane Granulated Sugar



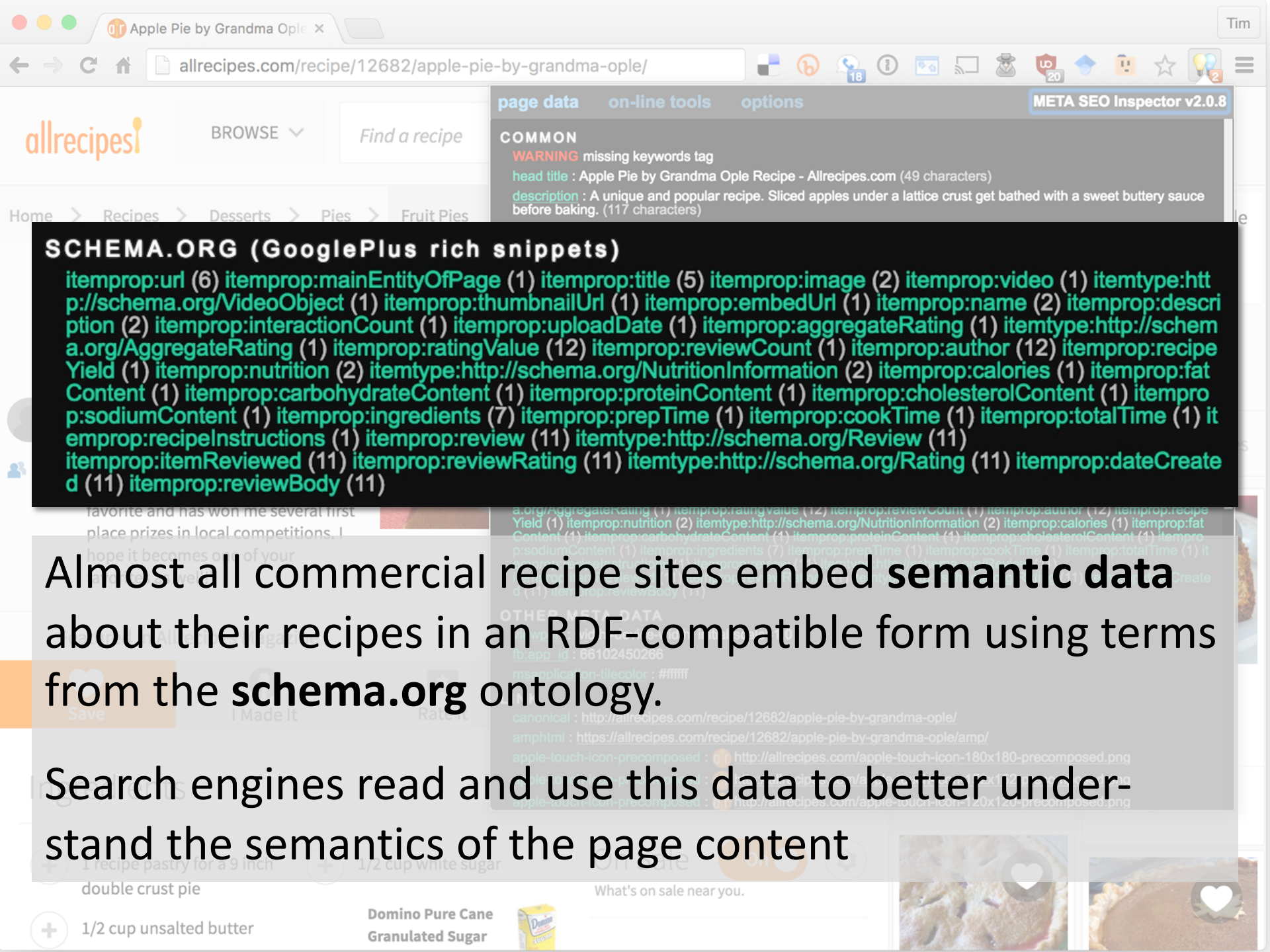
On Sale

On



What's on sale near you.





COMMON
WARNING missing keywords tag
head title : Apple Pie by Grandma Ople Recipe - Allrecipes.com (49 characters)
description : A unique and popular recipe. Sliced apples under a lattice crust get bathed with a sweet buttery sauce before baking. (117 characters)

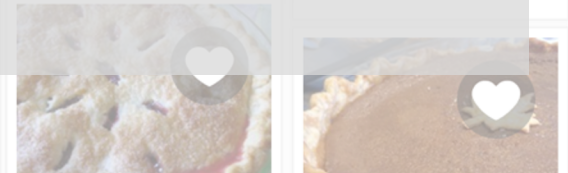
SCHEMA.ORG (GooglePlus rich snippets)
itemprop:url (6) itemprop:mainEntityOfPage (1) itemprop:title (5) itemprop:image (2) itemprop:video (1) itemtype:http://schema.org/VideoObject (1) itemprop:thumbnailUrl (1) itemprop:embedUrl (1) itemprop:name (2) itemprop:description (2) itemprop:interactionCount (1) itemprop:uploadDate (1) itemprop:aggregateRating (1) itemtype:http://schema.org/AggregateRating (1) itemprop:ratingValue (12) itemprop:reviewCount (1) itemprop:author (12) itemprop:recipeYield (1) itemprop:nutrition (2) itemtype:http://schema.org/NutritionInformation (2) itemprop:calories (1) itemprop:fatContent (1) itemprop:carbohydrateContent (1) itemprop:proteinContent (1) itemprop:cholesterolContent (1) itemprop:sodiumContent (1) itemprop:ingredients (7) itemprop:prepTime (1) itemprop:cookTime (1) itemprop:totalTime (1) itemprop:recipeInstructions (1) itemprop:review (11) itemtype:http://schema.org/Review (11) itemprop:itemReviewed (11) itemprop:reviewRating (11) itemtype:http://schema.org/Rating (11) itemprop:dateCreated (11) itemprop:reviewBody (11)

OTHER META DATA
canonical : http://allrecipes.com/recipe/12682/apple-pie-by-grandma-ople/
amphtml : https://allrecipes.com/recipe/12682/apple-pie-by-grandma-ople/amp/
apple-touch-icon-precomposed : http://allrecipes.com/apple-touch-icon-180x180-precomposed.png
apple-touch-icon-precomposed : http://allrecipes.com/apple-touch-icon-120x120-precomposed.png

Almost all commercial recipe sites embed **semantic data** about their recipes in an RDF-compatible form using terms from the **schema.org** ontology.

Search engines read and use this data to better understand the semantics of the page content

1/2 cup unsalted butter
Domino Pure Cane Granulated Sugar



Conversational Bots

Voice-driven conversational systems like Amazon Echo and Google Home use knowledge graphs to help understand our requests



Where does the knowledge come from?

- Initial knowledge graphs like *DBpedia* and *Freebase* started with data from **Wikipedia** and encoded it in custom ontologies
- Current focus is on extracting information from text of source documents, e.g., journal articles, Newswire, social media, etc.

NIST Text Analysis Conference



- Annual evaluation workshops since 2008 on natural language processing & related applications with large test collections and common evaluation procedures
- **Knowledge Base Population (KBP)** tracks focus on building KBs from information extracted from text
 - **Cold Start KBP:** construct a KB from text
 - **Entity discovery & linking:** cluster and link entity mentions
 - Slot filling
 - Slot filler validation
 - Sentiment
 - Events: discover and cluster events in text

<http://nist.gov/tac>

2016 TAC Cold Start KBP



- Read 90K documents: newswire articles & social media posts in English, Chinese and Spanish
- Find entity mentions, types and relations
- Cluster entities within and across documents and link to a reference KB when appropriate
- Remove errors (*Obama born in Illinois*), draw sound inferences (*Malia and Sasha sisters*)
- Create knowledge graph with provenance data for entities, mentions and relations

Information extraction from text



Identify relationships

Link concepts to entities

ebqids:hasMeans

http://dbpedia.org/resource/Buffer_overflow

ebqids:affectsProduct

CVE-2012-0150

Buffer overflow in msvcrt.dll in Microsoft Windows Vista SP2, Windows Server 2008 SP2, R2, and R2 SP1, and Windows 7 Gold and SP1 allows remote attackers to execute arbitrary code via a crafted media file, aka "Msvcrt.dll Buffer Overflow Vulnerability."

http://dbpedia.org/resource/Arbitrary_code_execution

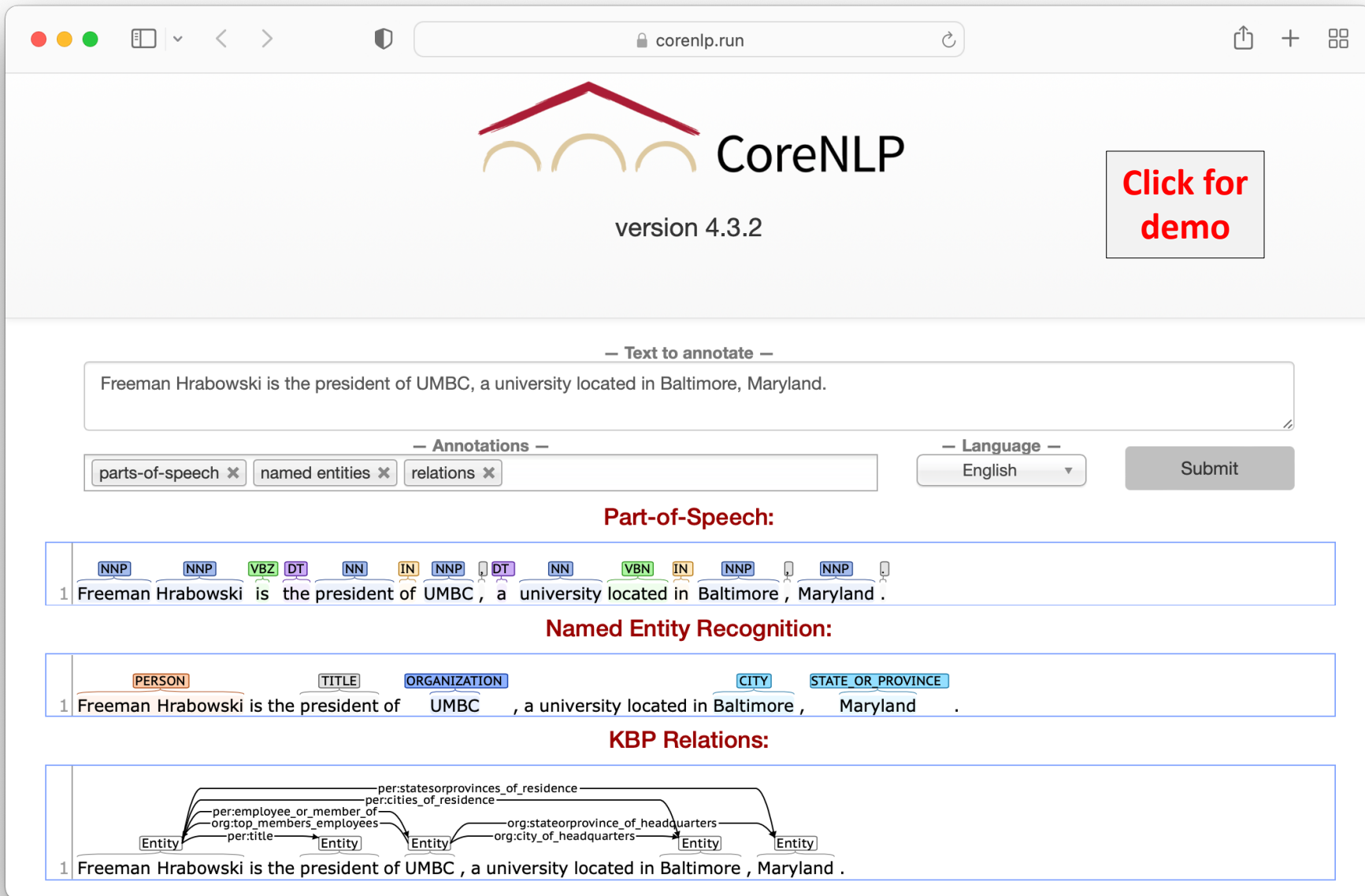
http://dbpedia.org/resource/Windows_7

- Information extraction techniques identify entities, relations and concepts in security related text
- Map to terms in our ontology and [DBpedia](#) knowledge graph
- Also map them to terms in the [Wikidata](#) knowledge graph

NLP Tools

- There is a rich and growing collection of open-source NLP tools
- Comprehensive pipelines:
 - Stanford CorNLP tools
 - Spacy
 - NLTK
- Word embeddings
 - Word2vec, BERT, Semsim

Stanford CoreNLP Tools



The screenshot displays the Stanford CoreNLP web interface. At the top, the browser address bar shows 'corenlp.run'. The page features the CoreNLP logo and version '4.3.2'. A 'Click for demo' button is visible on the right. The main interface includes a text input field containing the sentence: 'Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland.' Below the input, there are buttons for 'Annotations' (parts-of-speech, named entities, relations) and a 'Language' dropdown set to 'English'. A 'Submit' button is also present.

Part-of-Speech:

1. Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland .

Named Entity Recognition:

1. Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland .

KBP Relations:

1. Freeman Hrabowski is the president of UMBC, a university located in Baltimore, Maryland .

JSON/XML => KG triples

{ "text": "John Smith lives in Baltimore, Maryland. He is married to Mary Jones. She works at Loyola University where she is a professor. The university is in Baltimore.\n\n\n",

"docid": "text1.txt",

"corefs": {

"9": [

{ "endIndex": 6,

"animacy": "INANIMATE",

"text": "Baltimore",

"isRepresentativeMention": true,

"number": "SINGULAR",

"startIndex": 5,

"sentNum": 1,

"gender": "NEUTRAL",

"position": [1, 2q],

"headIndex": 5,

"type": "PROPER",

"id": 1

},

{

```
##### :e_text1_1 LOCATION "Baltimore" #####
```

```
:e_text1_1      type      LOCATION
```

```
:e_text1_1      canonical_mention "Baltimore"  text1:20-29
```

```
:e_text1_1      mention "Baltimore"  text1:20-29
```

```
:e_text1_1      mention "Baltimore"  text1:151-160
```

```
##### :e_text1_2 ORGANIZATION "Loyola University" #####
```

```
:e_text1_2      type      ORGANIZATION
```

```
:e_text1_2      canonical_mention "Loyola University"  text1:85-102
```

```
:e_text1_2      mention "Loyola University"  text1:85-102
```

```
:e_text1_2      mention "The university"  text1:130-144
```

```
##### :e_text1_3 PERSON "John Smith" #####
```

```
:e_text1_3      type      PERSON
```

```
:e_text1_3      canonical_mention "John Smith"  text1:0-10
```

```
:e_text1_3      mention "John Smith"  text1:0-10
```

```
:e_text1_3      mention "He"  text1:42-44
```

```
:e_text1_3      mention "She"  text1:72-75
```

```
:e_text1_3      mention "she"  text1:109-112
```

```
:e_text1_3      openie:lives_in :e_text1_1      text1:0-3
```

```
:e_text1_3      per:spouse      :e_text1_5      text1:42-43
```

```
:e_text1_3      openie:is_married_to :e_text1_5      text1:42-43
```

```
:e_text1_3      per:employee_of :e_text1_2      text1:72-74
```

Industrial-Strength Natural Language Processing

IN PYTHON

Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and tries to avoid wasting it. It's easy to install, and its API is simple and productive.

GET STARTED

Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If your application needs to process entire web dumps, spaCy is the library you want to be using.

FACTS & FIGURES

Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins, integrate with your machine learning stack and build custom components and workflows.

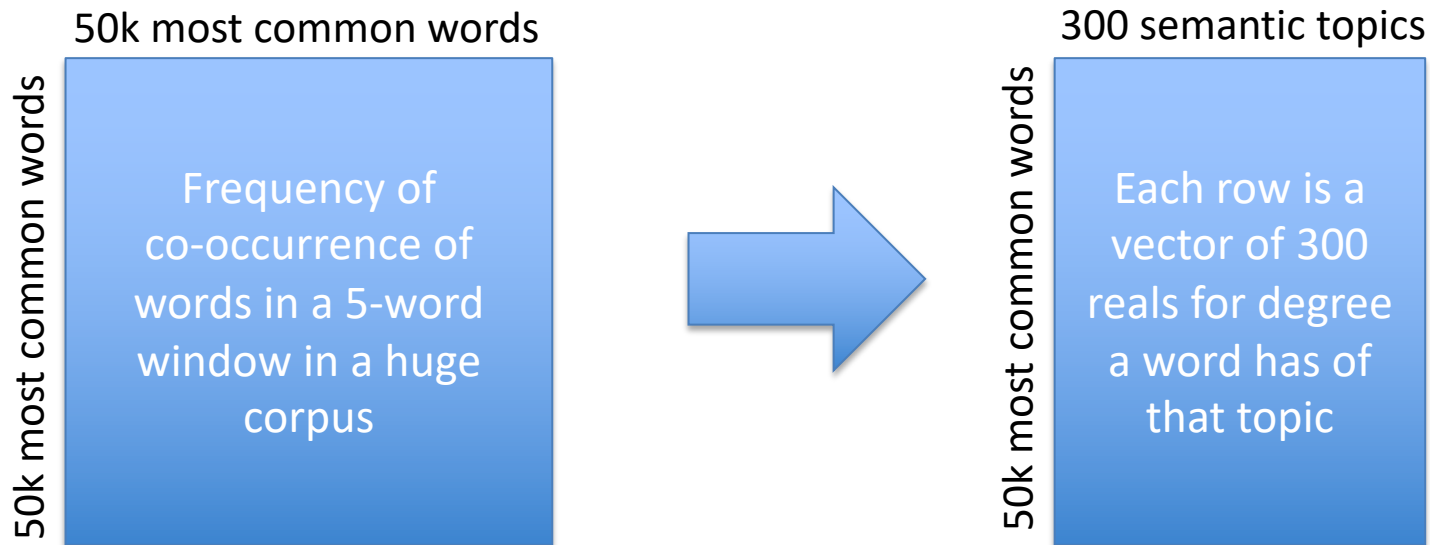
READ MORE

Learning word meaning?

- How can we learn what a word means?
- The linguist [John Rupert Firth](#) famously write in 1957
 - “You shall know a word by the company it keeps”
- A way to recognize that two words have similar meanings is to note that they occur in similar contexts
 - E.g., physician & doctor, nurse & doctor, love & hate

Word Embeddings

- [Latent Semantic Analysis](#) (LSA) learns a vector (e.g., 300 reals 0..1) for each unique word in a corpus to represent its meaning
 - LSA also used for document [topic modelling](#)
- An example of [dimensionality reduction](#)



Sentence similarity

How similar are the two sentences semantically on a scale of 0-5?

The mouse ate some cheese

Cheddar was eaten by a rat



3.824

Pearson's Correlation



Close enough!

It's a 4!



We used this approach in 2013 to win in a sentence similarity task

UMBC semantic similarity service

UMBC Semantic Similarity Serv x +

← → ↻ 🏠 ⓘ Not Secure | swoogle.umbc.edu/SimS... 🔍 ☆ G ⓘ 🚫 🇺🇸 🗨️ 🤖 📶 🔊 ⋮

UMBC Top-N Similarity Service

[Go back](#)

The input word:

Part of Speech: Noun Verb Adjective Adverb

The value of N: 10 20 30 40 50 100

Type: Concept Similarity Relation Similarity

Corpus: Refined Stanford WebBase corpus LDC English Gigawords Corpus (American newswire services only)

word2vec

WORD2VEC

Uses a shallow neural network to map words to a vector space where words with similar contexts have close vectors.



Chris Albon

Word2Vec

- Developed by Google also in 2013 using a neural network approach
- Two models: CBOW and skip grams
- Trained on a much larger corpus from the Web
- Models can be downloaded and are still used today
 - E.g., the spaCy NLP system includes word2vec to measure similarity

Word2vec demo

Click for demo

Models

Select one of the available models

English GoogleNews Negative300

Nearest words

Given a word, this demo shows a list of other words that are similar to it, i.e. nearby in the vector space.

Type in a word **Show nearest** Case sensitive: Top N: 10

Similarity of two words

Given two words, this demo gives the similarity value between 1 and -1.

Type in a word Type in a word **Show similarity**

Word analogy

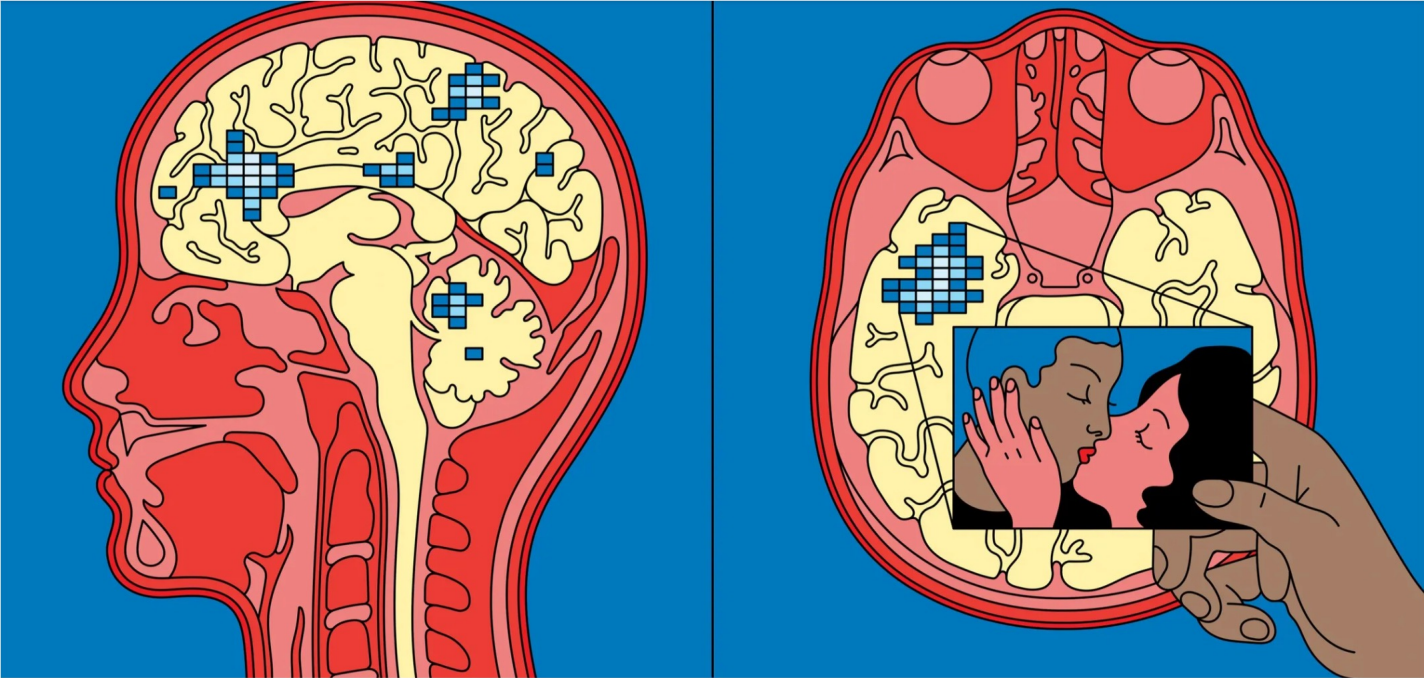
This demo computes word analogy: the first word is to the second word like the third word is to which word? Try for example *ilma - lintu - vesi* (air - bird - water) which would expect to return *kala* (fish) because fish is to water like birds is to air. Other cases could be for example *sammakko - hyppää - kala*. This is however only a toy to show what is possible - most of the time the analogy does not work particularly well (at least for the Finnish data).

Type in a word Type in a word Type in a word **Show** Top N: 2

Scientists using fMRI to measure brain activity find locations associated with similar concepts – brain embeddings!

THE NEW YORKER Newsletter My A

SCIENCE AND TECH [The Darker Side of Aaron Swartz](#) [The Friendship That Made Google Huge](#) [Bones of Contention](#) [The Histories Hidden in the Periodic Table](#)



It isn't so much that brain scans have improved—it's that we've got better at reading them. Illustration by Laura Edelbacher

ANNALS OF TECHNOLOGY DECEMBER 6, 2021 ISSUE

THE SCIENCE OF MIND READING

Researchers are pursuing age-old questions about the nature of thoughts—and learning how to read them.

By James Somers
November 29, 2021

**Click to
read**

Conclusion



- KGs help in extracting information from text
- The information extracted can update the KGs
- The KGs provide support for new tasks, such as question answering, speech interfaces and produce data useful in applications, like IDSs
- Their use will grow and evolve in the future
- New machine learning frameworks will result in better accuracy