— 26 March 2018 —

SERTIS

CONNECTING DATA TO ACTION

Natural Language Processing in Industry

Jussi Jousimo, PhD

Senior Data Scientist @ Sertis

Outline

- Introduction
- NLP applications in industry
- Some NLP applications in more detail
- Practical implementation of NLP applications
- Bottom line
- NLP at Sertis

Introduction

Why NLP?

44 zettabytes (44 trillion GB) of data in the world 2020

* IDC, EMC

70-80% unstructured data including text and voice

* Merrill Lynch 1998

Total AI market $16 billion in 2017

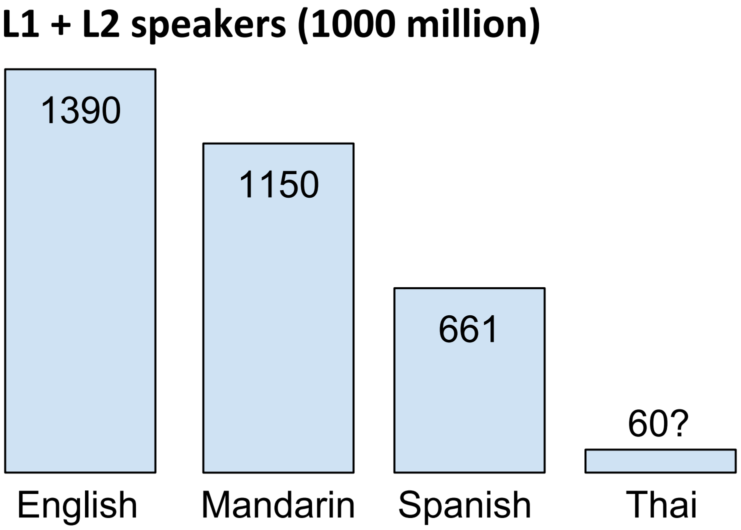
NLP market $16 billion in 2021

Total AI market $191 billion in 2025

* MarketsAndMarkets

Natural Language Processing Market worth 16.07 Billion USD by 2021: https://www.marketsandmarkets.com/PressReleases/natural-language-processing-nlp.asp
Artificial Intelligence Market by Offering […] https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-market-74851580.html

Natural Languages

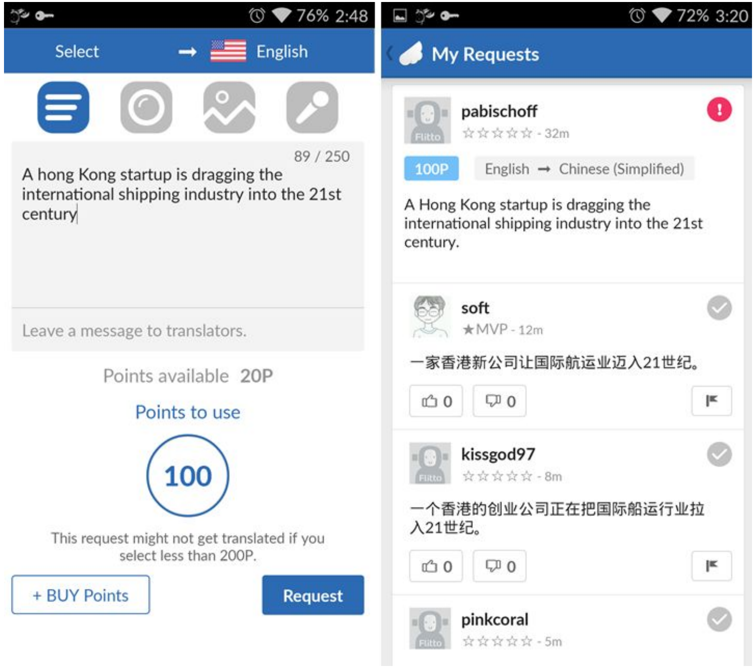


* SIL International 2017

List of languages by total number of speakers: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers
Languages used on the Internet: https://en.wikipedia.org/wiki/Languages_used_on_the_Internet#Content_languages_for_websites

NLP R&D in Industry

- English most researched, most applications
- Research effort not directly proportional to number of speakers
  - ○ How much research on specific properties of language?
  - ○ How much available training data?
- Data collection and labelling often expensive
  - ○ Smarter data collection, e.g. Flitto
  - ○ Data itself as a business model



89 / 250

A hong Kong startup is dragging the international shipping industry into the 21st century

Leave a message to translators.

Points available 20P

Points to use

100

This request might not get translated if you select less than 200P.

pabischoff ☆☆☆☆☆ - 32m

100P English → Chinese (Simplified)

A Hong Kong startup is dragging the international shipping industry into the 21st century.

soft ★MVP - 12m

一家香港新公司让国际航运业迈入21世纪。

kissgod97 ☆☆☆☆☆ - 8m

一个香港的创业公司正在把国际船运行业拉入21世纪。

pinkcoral ☆☆☆☆☆ - 5m

Smarter data collection with Flitto

Korean entrepreneur went from translating K-Pop tweets to selling language data to web giants: https://www.techinasia.com/korean-entrepreneur-translating-kpop-tweets-selling-language-data-google

Examples of NLP Applications

Topic classification	Topic clustering	Tagging	Sentiment analysis
Aboutness	Summarization	Search	Document similarity
Machine translation	Chatbots	Q&A	Speech recognition
Speech-to-text	Text-to-speech	Spell correction	etc.

Industries

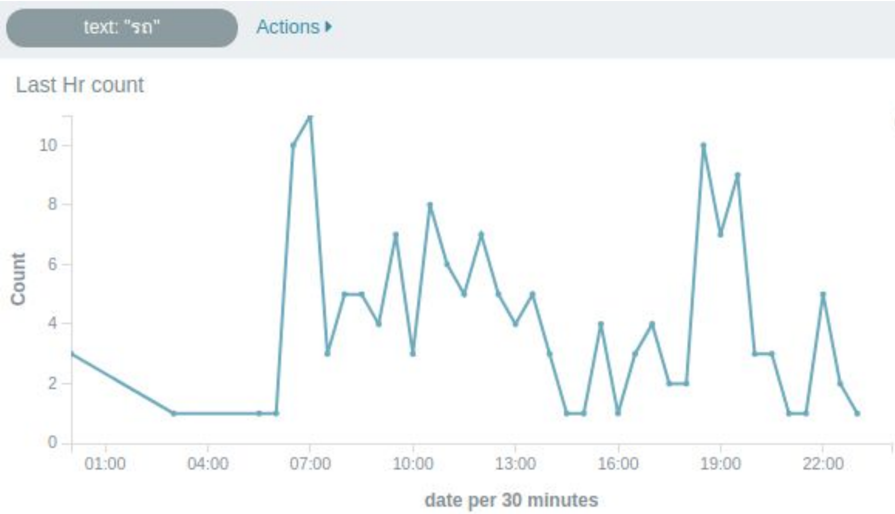
Retail

- **Product search**
  - 67% increase in conversion vs. site average * Econsultancy 2013
  - Most queries include 1-3 words * SLI Systems 2017
  - Text or voice, e.g. Amazon Alexa
  - Additional features: advanced search, semantic search, autocompletion, recommendations
- **Product description text mining**
  - Identify entities for improved search
  - Enhance recommendations
- **Engagement / seller chatbots**
- **Partially automated customer service**

Is site search less important for niche retailers?: https://econsultancy.com/blog/62401-is-site-search-less-important-for-niche-retailers
Natural Language Processing and eBay Listings: https://www.ebayinc.com/stories/blogs/tech/natural-language-processing-and-ebay-listings/

Media & Marketing

- Follow **market trends**, brands, companies, persons, events, locations, ...
- Marketing **campaign monitoring** from news, social media
- **Sentiment analysis** on markets, product / service reviews, ...

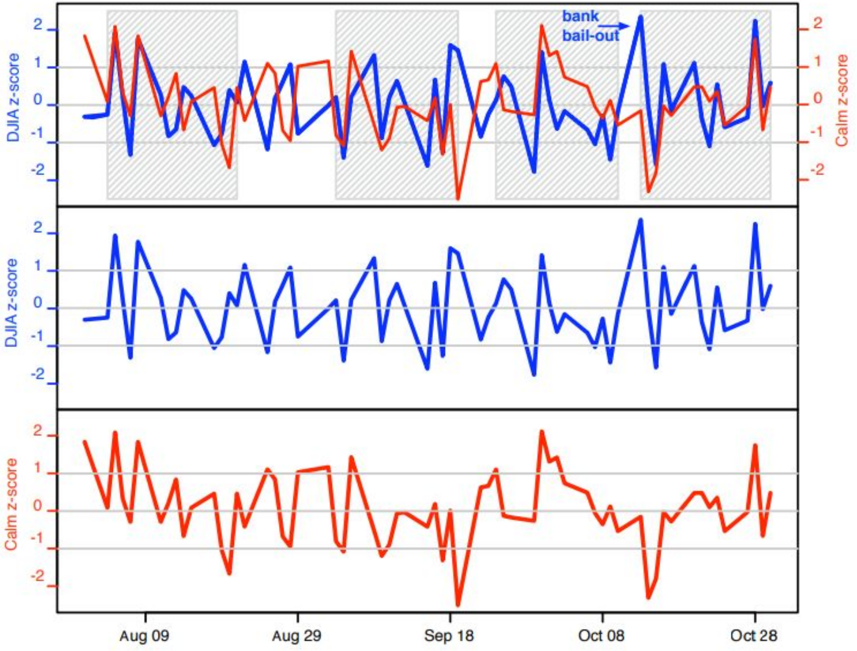


© Sertis Co. Ltd.

Do traffic jams in Bangkok correlate with mentions of cars from Tweets?

Investing & Finance & Insurance

- **Sentiment analysis** on news, SEC filings, social media
- **Early mentions**
- **Enrich** financial news
  - Entity recognition: people, organizations, places, events, ...
  - Topic classification
  - E.g. Thomson Reuters' Open Calais
- **Summarization**
- **Compliance, fraud detection**

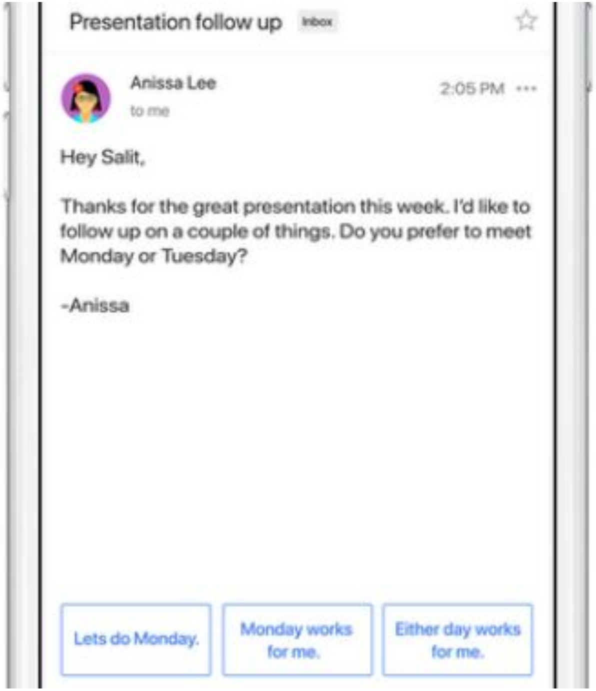


Bollen et al. 2010 - Twitter mood predicts the stock market: https://arxiv.org/abs/1010.3003

Open Calais: http://www.opencalais.com/

Messaging & Social Media

- **Spam and fake news detection**
- **Intention recognition**
  - E.g. detect sales ⇒ provide sales tools
- Content **recommendations**
- Personalized **advertising**
- **Voice typing**
- **Next word prediction**
- E.g. Facebook’s DeepText
- E.g. Google’s Smart Reply



Presentation follow up

Anissa Lee
to me
2:05 PM

Hey Salit,

Thanks for the great presentation this week. I'd like to follow up on a couple of things. Do you prefer to meet Monday or Tuesday?

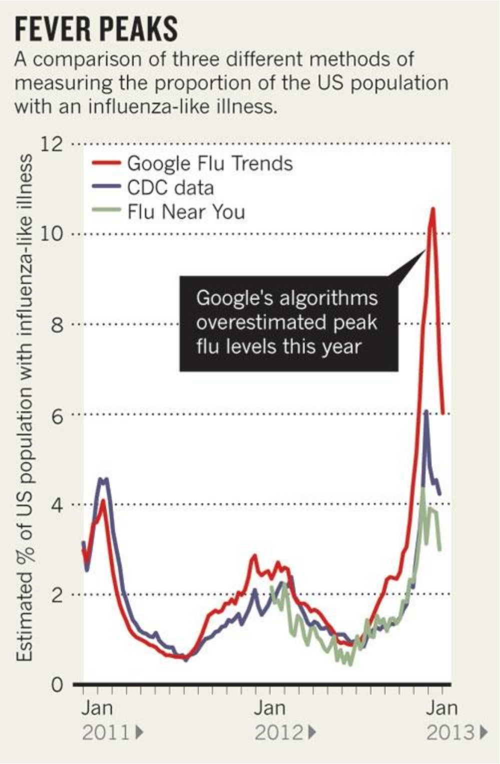
-Anissa

Google’s Smart Reply

Introducing DeepText: Facebook's text understanding engine:
https://code.facebook.com/posts/181565595577955/introducing-deeptext-facebook-s-text-understanding-engine/
Efficient Smart Reply, now for Gmail:
https://research.googleblog.com/2017/05/efficient-smart-reply-now-for-gmail.html

Healthcare

- **Transcription**, **annotation** and **summarization** of clinical notes, medical journals, publications
- **Semantic search** for clinical questions from medical notes
    - E.g. CogStack project
- **Diagnosis support** (from description of symptoms, speech analysis)
- E.g. IBM’s Watson Health, Woebot, Google Flu Trends



FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.

Butler - When Google got flu wrong (Nature 2013): https://www.nature.com/news/when-google-got-flu-wrong-1.12413

Artificial intelligence in health care: within touching distance: http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)31540-4/fulltext
Welcome to Cogstack: https://cogstack.github.io/
IBM Watson Health: https://www.ibm.com/watson/health/
Woebot: https://woebot.io/
Google Flu Trends: https://en.wikipedia.org/wiki/Google_Flu_Trends

Education & Research

- **Language learning**
  - Correct written and speaking errors
  - Teaching chatbot
  - Personalized teaching
  - Translations
- **Plagiarism detection**
  - Issues with rephrased or translated texts

Customer Service

- **Chatbots**
  - $4.5 billion annual cost saving by 2022 * Juniper 2018
  - 9% of Fortune 500 companies work with chatbots * Forrester Research 2017
  - Orders, bookings, service requests, feedback or other focused scope
  - High volume of requests
  - Multiple variables
  - Routing to human agents
- **Speaker recognition / separation from call center calls**
  - Speech-to-text
  - Issue classification

Chatbot Conversations to deliver $8 billion in Cost savings by 2022: https://www.juniperresearch.com/analystxpress/july-2017/chatbot-conversations-to-deliver-8bn-cost-saving
Chatbot Commerce Benchmark: https://www.forrester.com/report/Chatbot+Commerce+Benchmark/-/E-RES137221
Chatbots in customer service: https://www.accenture.com/t00010101T000000__w__/br-pt/_acnmedia/PDF-45/Accenture-Chatbots-Customer-Service.pdf

Other

- **Recruiting**
  - Information extraction from CVs, job descriptions
- **News**
  - Natural language generation from structured data (weather, sports, finance)
- **Legal**
  - Semantic search of legal documents, laws
  - Structurize, classify and link legal documents and legislation
  - Summarization
- **Other**
  - Spelling correction, e.g. Microsoft’s Word
  - OCR error correction

Applications in Detail

Search: Introduction

Difficult problem

○ Provide relevant results but what is relevant?

○ What is user’s intention and context?

○ Fast

○ Easy to use

Indexing

Store and index documents in the way that it enables quick search

Querying

Interpret search query and filter documents matching it

Ranking

Sort documents given relevancy to the user intent and context

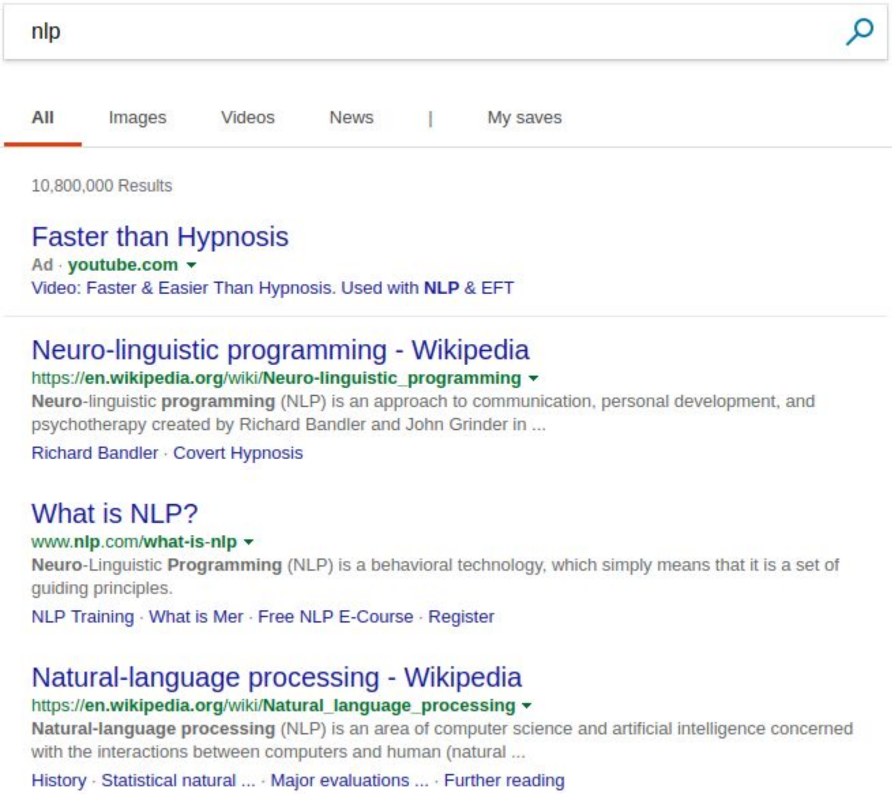
Search: Approaches

Keyword based

- NLP pipeline for documents and queries
- Inverted index
- Rank by term/document frequencies

Semantic

- Intent (word co-occurrences and distances, concepts)
- Context (location, query history, trends)
- Synonyms
- Relationship of entities (knowledge graph)
- Natural language interface
- Related queries



nlp

All · Images · Videos · News · | · My saves

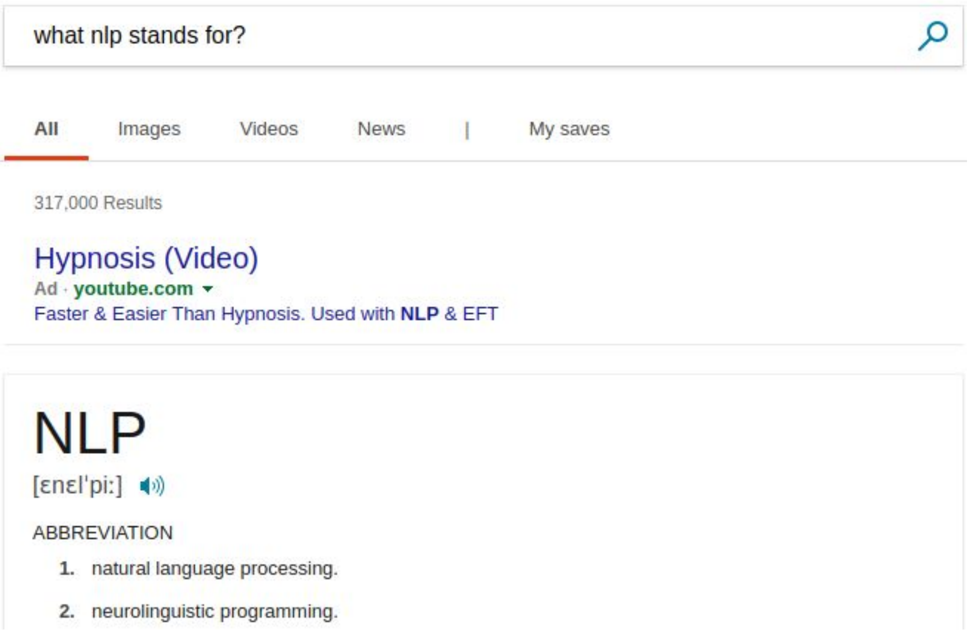
10,800,000 Results

Faster than Hypnosis
Ad · youtube.com ▾
Video: Faster & Easier Than Hypnosis. Used with **NLP** & EFT

Neuro-linguistic programming - Wikipedia
https://**en.wikipedia.org**/wiki/**Neuro-linguistic_programming** ▾
Neuro-linguistic **programming** (NLP) is an approach to communication, personal development, and psychotherapy created by Richard Bandler and John Grinder in ...
Richard Bandler · Covert Hypnosis

What is NLP?
www.**nlp**.com/**what-is-nlp** ▾
Neuro-Linguistic **Programming** (NLP) is a behavioral technology, which simply means that it is a set of guiding principles.
NLP Training · What is Mer · Free NLP E-Course · Register

Natural-language processing - Wikipedia
https://**en.wikipedia.org**/wiki/**Natural_language_processing** ▾
Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural ...
History · Statistical natural ... · Major evaluations ... · Further reading



what nlp stands for?

All · Images · Videos · News · | · My saves

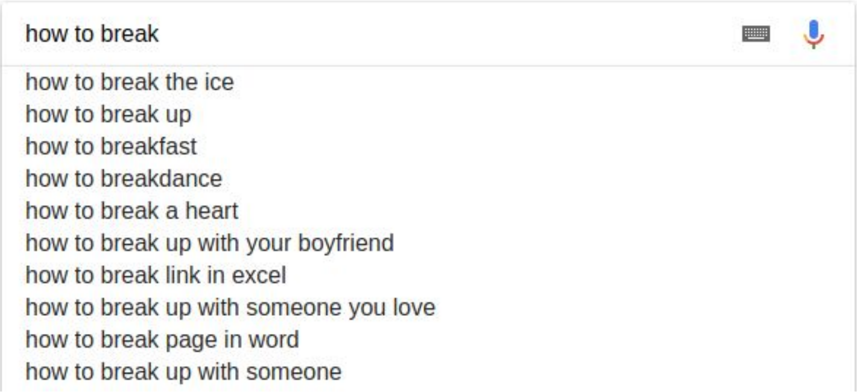
317,000 Results

Hypnosis (Video)
Ad · youtube.com ▾
Faster & Easier Than Hypnosis. Used with **NLP** & EFT

NLP
[ɛnɛl'piː] 🔊

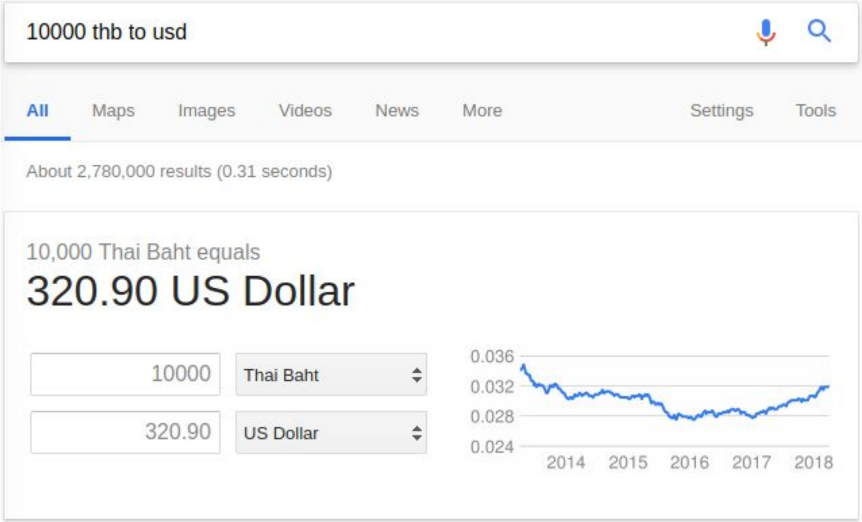
ABBREVIATION

1. natural language processing.
2. neurolinguistic programming.



how to break

how to break the ice
how to break up
how to breakfast
how to breakdance
how to break a heart
how to break up with your boyfriend
how to break link in excel
how to break up with someone you love
how to break page in word
how to break up with someone



10000 thb to usd

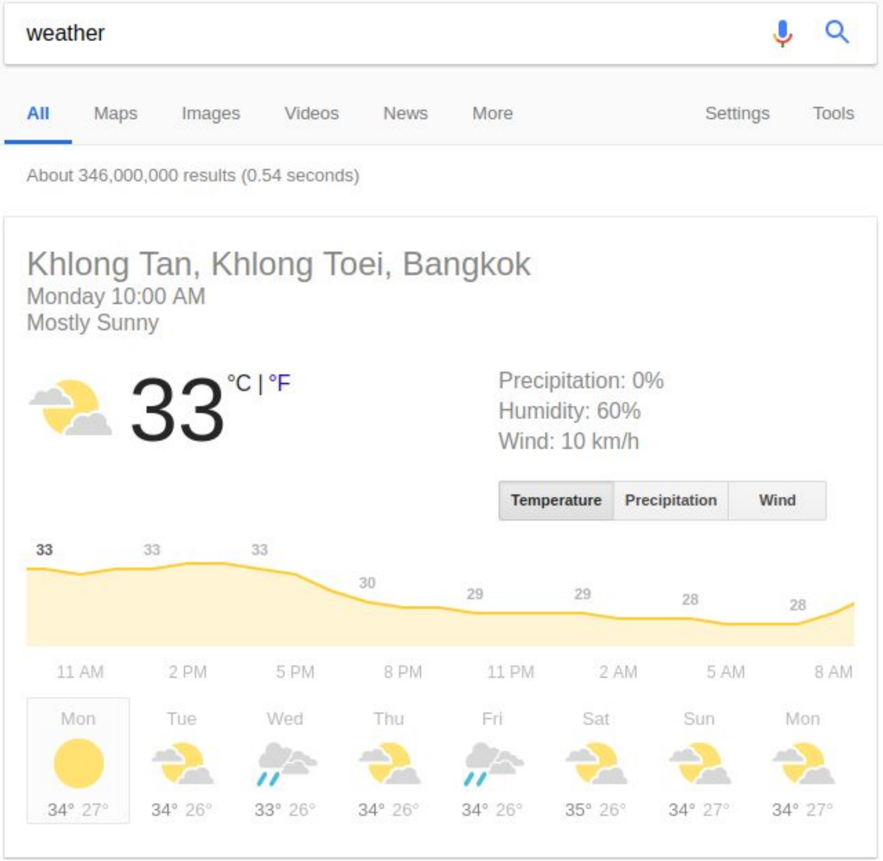
About 2,780,000 results (0.31 seconds)

10,000 Thai Baht equals

320.90 US Dollar

10000 Thai Baht

320.90 US Dollar



weather

About 346,000,000 results (0.54 seconds)

Khlong Tan, Khlong Toei, Bangkok
Monday 10:00 AM
Mostly Sunny

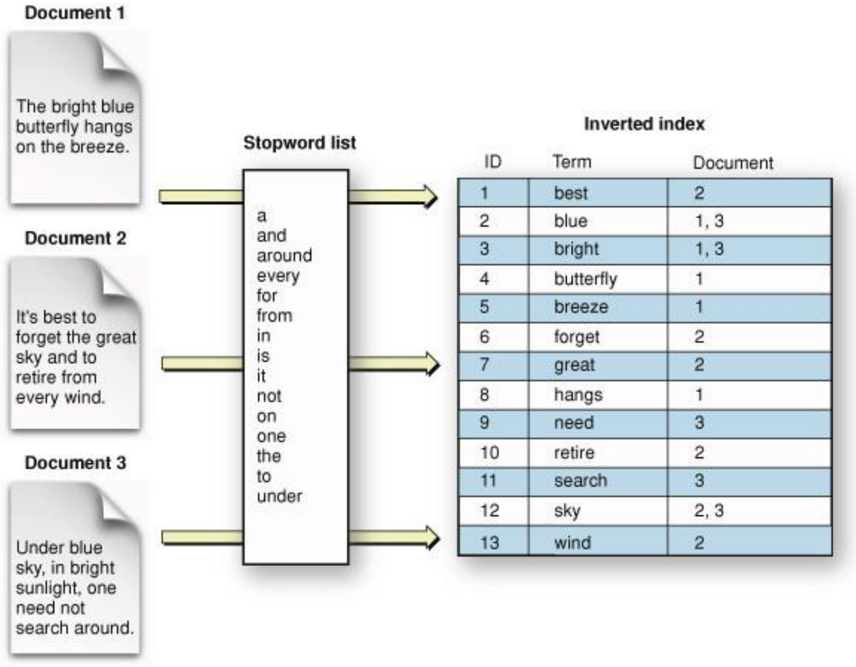
33 °C | °F

Precipitation: 0%
Humidity: 60%
Wind: 10 km/h

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon
34° 27°	34° 26°	33° 26°	34° 26°	34° 26°	35° 26°	34° 27°	34° 27°

Search: Keyword Based

- Inverse index
  - Document ingestion pipeline: extract text, tokenize, normalize, stop words removal, NER, …
  - Index: term ⇒ document
  - Fast search
- Querying
  - Boolean algebra, proximity matching, partial matching, (negative) boosting
- Ranking
  - Boolean model, TF/IDF, cosine similarity, field length
- Implemented e.g. in Apache Solr and ElasticSearch



Document 1

The bright blue butterfly hangs on the breeze.

Document 2

It's best to forget the great sky and to retire from every wind.

Document 3

Under blue sky, in bright sunlight, one need not search around.

Stopword list

a
and
around
every
for
from
in
is
it
not
on
one
the
to
under

Inverted index

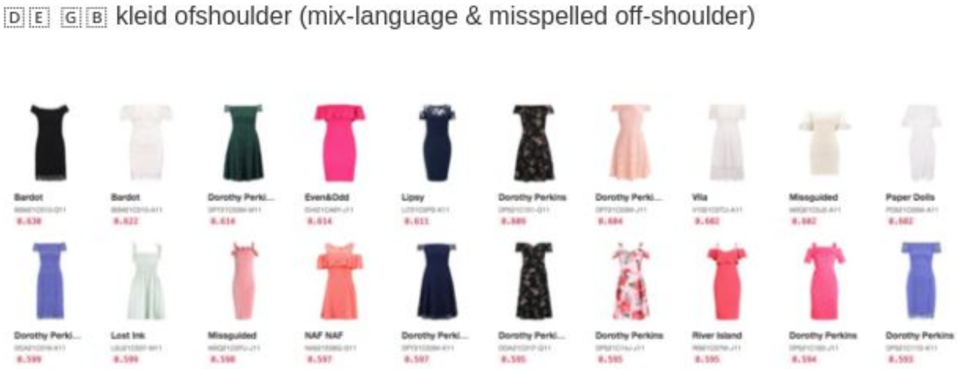
ID	Term	Document
1	best	2
2	blue	1, 3
3	bright	1, 3
4	butterfly	1
5	breeze	1
6	forget	2
7	great	2
8	hangs	1
9	need	3
10	retire	2
11	search	3
12	sky	2, 3
13	wind	2

Dealing with Human Language:
https://www.elastic.co/guide/en/elasticsearch/guide/current/languages.html
Search in Depth
https://www.elastic.co/guide/en/elasticsearch/guide/current/search-in-depth.html
Apache Solr: http://lucene.apache.org/solr/

Apple Developer - Search Basics
https://developer.apple.com/library/content/documentation/UserExperience/Conceptual/SearchKitConcepts/searchKit_basics/searchKit_basics.html

Search: Deep Learning

- Map queries with clicks
  - Spelling mistakes, multiple languages, synonyms, entities, ambiguity, etc. in the same model
  - Requires query/click data or proactive approach
  - Combine with image and other data (multimodal learning)
- 15% of Google’s queries are new each day



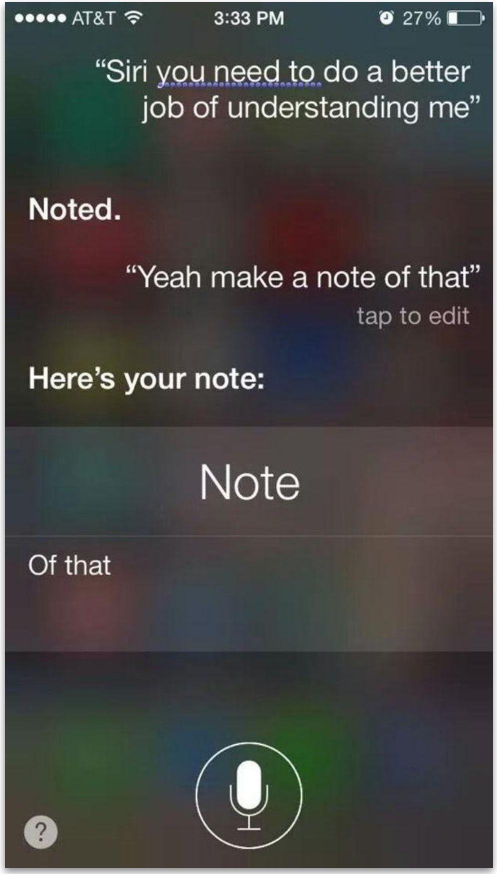
D E G B kleid ofshoulder (mix-language & misspelled off-shoulder)

Han Xio - Cross-Lingual End-to-End Product Search with Deep Learning: https://jobs.zalando.com/tech/blog/search-deep-neural-network/

Deep Learning for Search: https://www.manning.com/books/deep-learning-for-search

Dialog Systems

- General and task-specific chatbots
- General chatbot and long dialogs difficult, partially because of too high expectations
- Higher success in narrow domains and with short dialogs
- Architectures:
  - Rule based
  - Corpus based: information retrieval, transduction
  - Frame based: gather information from user to fill frame
- Hybrid chatbot with human intervention
- Mostly mass products



"Siri you need to do a better job of understanding me"

Noted.

"Yeah make a note of that"

tap to edit

Here's your note:

Note

Of that

Jurafsky et al. 2017 - Dialog Systems and Chatbots: https://web.stanford.edu/~jurafsky/slp3/29.pdf

Speech Recognition

- Efficient way of communicating
- Human level recognition reached in controlled environment (2017)
- Next
  - Background noise, different accents, multiple speakers
  - Mixed languages
- Used for
  - Voice search, e.g. Google, Amazon
  - Personal assistants, e.g. Apple’s Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Baidu Duer, Samsung Bixby
  - Speech-to-text
  - Speech analytics
  - Identification

Personal assistant devices

Microsoft researchers achieve new conversational speech recognition milestone: https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/

Question Answering

- Questions
  - Specific domain
  - Open domain
- Architectures
  - Information retrieval based – extract answer from text documents
  - Knowledge based – extract answer from structured data
  - Multi-source – multiple data sources
- IBM’s DeepQA beat human opponent in Jeopardy! in 2011
- Commercial applications
  - Wolfram Alpha
  - Google Search
  - Microsoft Bing
  - Apple’s Siri

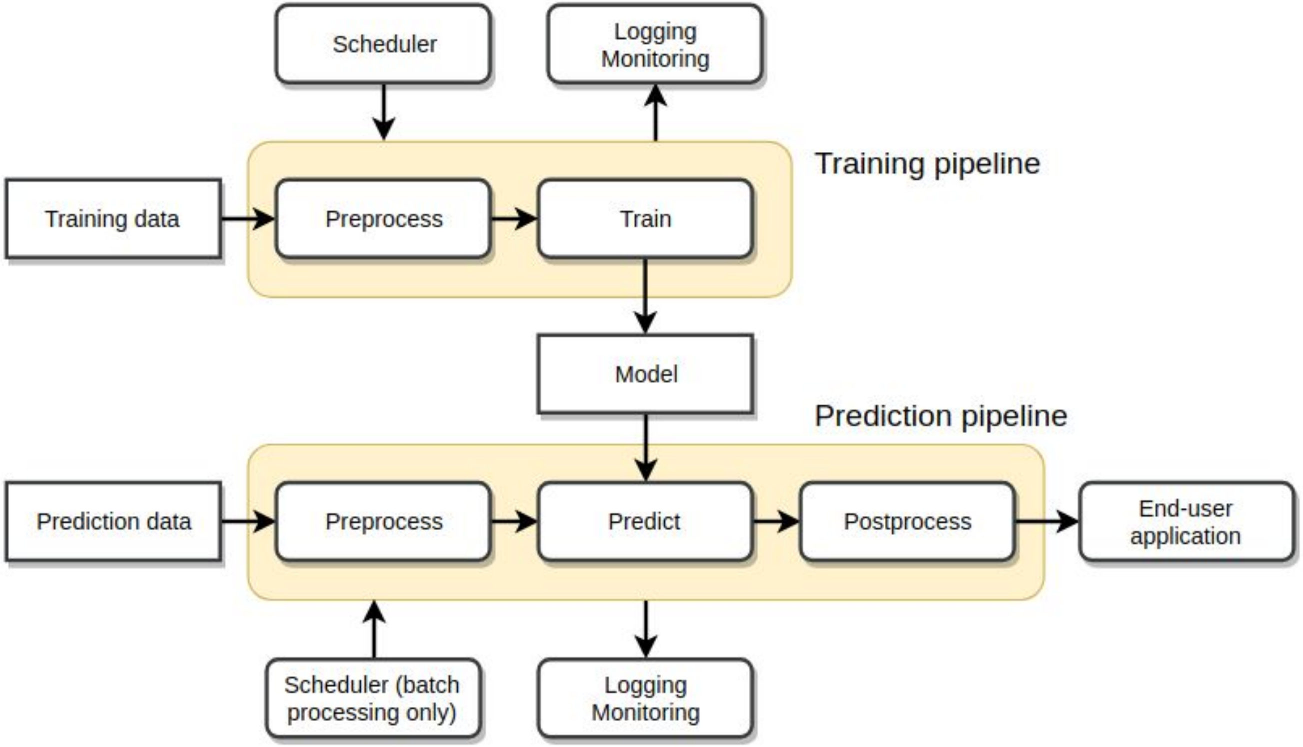
Jurafsky et al. 2017 - Question Answering https://web.stanford.edu/~jurafsky/slp3/28.pdf
IBM Watson: https://en.wikipedia.org/wiki/Watson_(computer)

Practical Implementation of NLP Applications

Implementation Considerations

<table>
<tr><td>Accuracy</td><td>Speed</td><td>Scaling</td><td></td></tr>
<tr><td colspan="2">Choice of frameworks</td><td colspan="2">Choice of cloud provider or custom</td></tr>
<tr><td colspan="2">Edge computing</td><td colspan="2">Integration with existing systems</td></tr>
<tr><td colspan="2">Logging and monitoring</td><td colspan="2">Maintenance</td></tr>
<tr><td colspan="4">Orchestration between data scientists, data engineers, software developers, etc.</td></tr>
</table>

Framework for ML/NLP Applications



NLP as a Service

English, Chinese and some other major languages provided by multiple vendors

Sentiment ⚘	POS tagging ⚘	Entity recognition	Topic classification
Aboutness	Language detection	Machine translation ⚐	Chatbots ⚐
Speech-to-text ⚐	Semantic search	Tokenization ⚘	

Thai support: ⚐ Google ⚘ NECTEC

Thai NLP Libraries

- PyThaiNLP (Python)
- Apache Lucene (Java)
- Facebook’s fastText
- Polyglot (Python)

PyThaiNLP: https://github.com/PyThaiNLP/
fastText: https://research.fb.com/fasttext/
Thai Tokenizer: https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-thai-tokenizer.html
Polyglot: https://github.com/aboSamoor/polyglot
Thai Natural Language Processing (Thai NLP) Resource: https://github.com/kobkrit/nlp_thai_resources

Challenges

- NLP still “academic”
- Context dependency, ambiguity, dialects, difficult to generalize across domains, ...
- **Scarce model training resources**
  - State-of-the art approach (deep learning) require lots of data
  - Pronounced for low resource languages such as Thai
- Natural languages evolve constantly
  - New words
  - New concepts
  - Social media text
  - “ออเจ้า”
- Sophisticated models not necessarily suitable for production
- Expectations too high
- Skilled NLP/ML/DE practitioners hard to find

Future

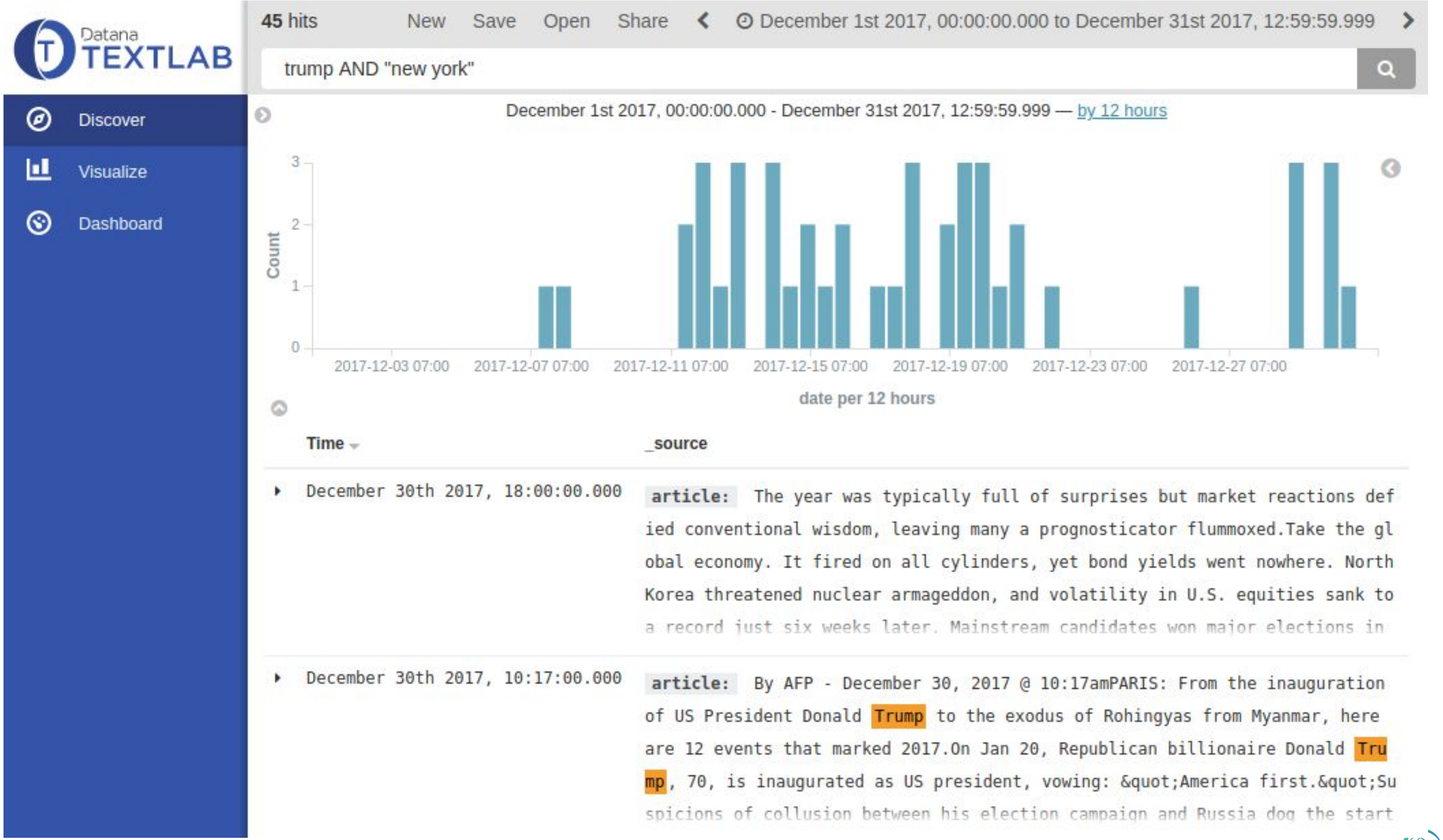
- Improved NLP applications in narrow and general domains
- More NLP enabled “smart” applications / services
- More industries using NLP
- Catch up with computer vision
  - NASNet, text generation, discrete sequence GANs, text style transfer, unsupervised methods, …
- Towards general search engines / chatbots / virtual assistants
  - May eventually merge

Learning Transferable Architectures for Scalable Image Recognition: https://arxiv.org/pdf/1707.07012.pdf

Bottom Line

- Know the business problems to solve
- Can machine learning solve the problems in practice
- Return of investment
- Evaluate service
- Fancy models might be too complex in production
- Models only a small part of production systems

NLP at Sertis



45 hits

December 1st 2017, 00:00:00.000 to December 31st 2017, 12:59:59.999

trump AND "new york"

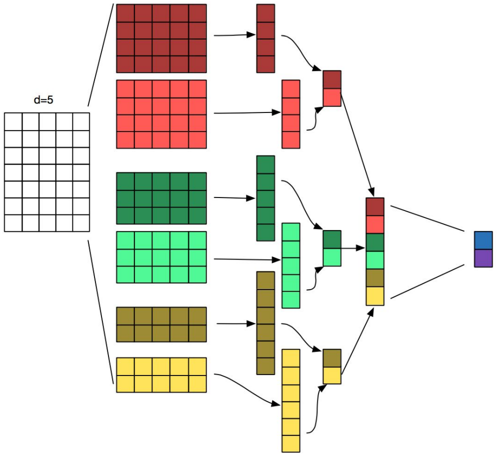
December 1st 2017, 00:00:00.000 - December 31st 2017, 12:59:59.999 — by 12 hours

Time	_source
December 30th 2017, 18:00:00.000	**article:** The year was typically full of surprises but market reactions defied conventional wisdom, leaving many a prognosticator flummoxed.Take the global economy. It fired on all cylinders, yet bond yields went nowhere. North Korea threatened nuclear armageddon, and volatility in U.S. equities sank to a record just six weeks later. Mainstream candidates won major elections in
December 30th 2017, 10:17:00.000	**article:** By AFP - December 30, 2017 @ 10:17amPARIS: From the inauguration of US President Donald Trump to the exodus of Rohingyas from Myanmar, here are 12 events that marked 2017.On Jan 20, Republican billionaire Donald Trump, 70, is inaugurated as US president, vowing: &quot;America first.&quot;Suspicions of collusion between his election campaign and Russia dog the start

Topic Classification

us stocks futures flat digest record run and before datum blast future dow up _NUM_ pt s&p down _NUM_ pt nasdaq down _NUM_ pt by yashaswini swamynathan feb _NUM_ reuters u.s. stock index future be little change on wednesday ahead of a blast of economic datum and a day after federal reserve chair janet yellen paint a largely upbeat picture of the economy yellen say on tuesday before the u.s. senate banking committee that delay interest rate hike would be unwise but do not indicate when the fed would raise rate -PRON- testimony ...

germany say will accompany opel psa tie up talk berlin feb _NUM_ reuters the german government say on wednesday -PRON- would accompany talk on peugeot maker psa \'s < peup.pa > plan to buy general motors < gm.n > european business opel and that -PRON- have a strong interest in opel \'s future the government have a strong interest in a successful future for the business and -PRON- site of course this be about corporate decision and -PRON- have no evaluation to give on that government spokesman...

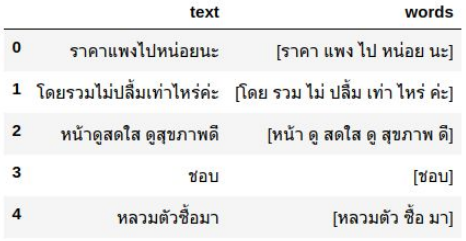


http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

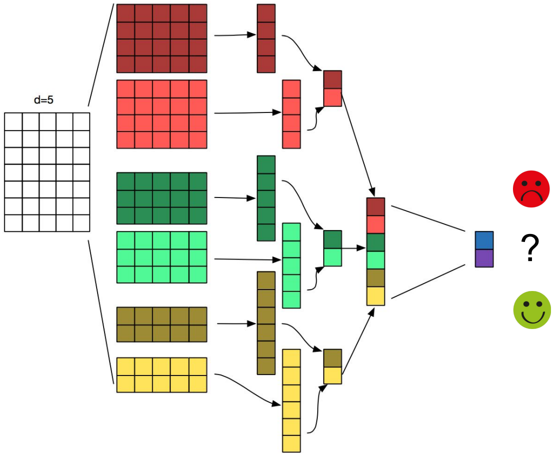
Economy
Finance / investing
...

Automotive
Energy
...

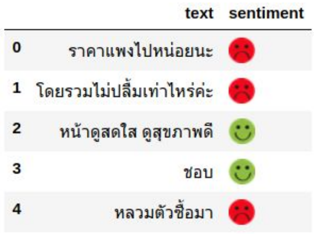
Sentiment Analysis



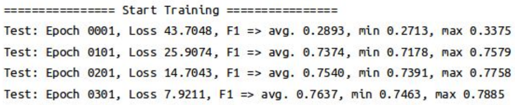
	text	words
0	ราคาแพงไปหน่อยนะ	[ราคา แพง ไป หน่อย นะ]
1	โดยรวมไม่ปลื้มเท่าไหร่ค่ะ	[โดย รวม ไม่ ปลื้ม เท่า ไหร่ ค่ะ]
2	หน้าดูสดใส ดูสุขภาพดี	[หน้า ดู สดใส ดู สุขภาพ ดี]
3	ชอบ	[ชอบ]
4	หลวมตัวซื้อมา	[หลวมตัว ซื้อ มา]



http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/



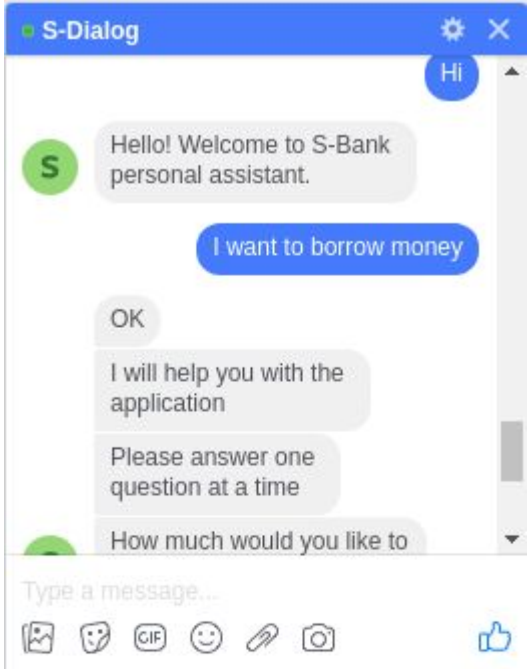
	text	sentiment
0	ราคาแพงไปหน่อยนะ	🙁
1	โดยรวมไม่ปลื้มเท่าไหร่ค่ะ	🙁
2	หน้าดูสดใส ดูสุขภาพดี	🙂
3	ชอบ	🙂
4	หลวมตัวซื้อมา	🙁



```
================ Start Training ================
Test: Epoch 0001, Loss 43.7048, F1 => avg. 0.2893, min 0.2713, max 0.3375
Test: Epoch 0101, Loss 25.9074, F1 => avg. 0.7374, min 0.7178, max 0.7579
Test: Epoch 0201, Loss 14.7043, F1 => avg. 0.7540, min 0.7391, max 0.7758
Test: Epoch 0301, Loss 7.9211, F1 => avg. 0.7637, min 0.7463, max 0.7885
```

⋮

Chatbots



Hi

S

Hello! Welcome to S-Bank personal assistant.

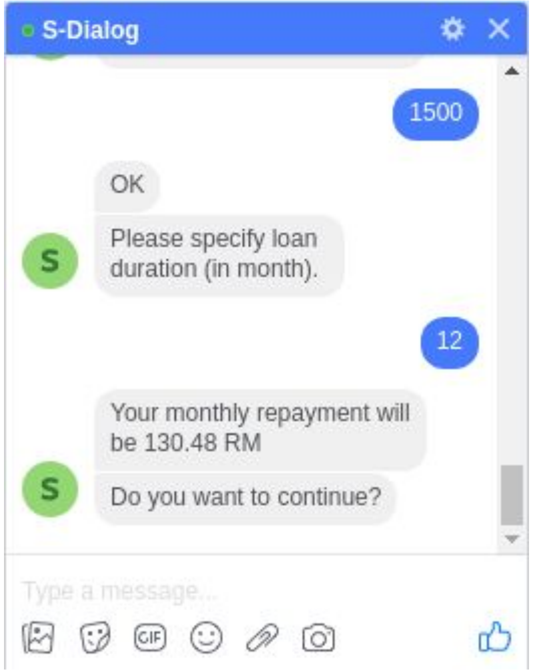
I want to borrow money

OK

I will help you with the application

Please answer one question at a time

How much would you like to



1500

OK

S

Please specify loan duration (in month).

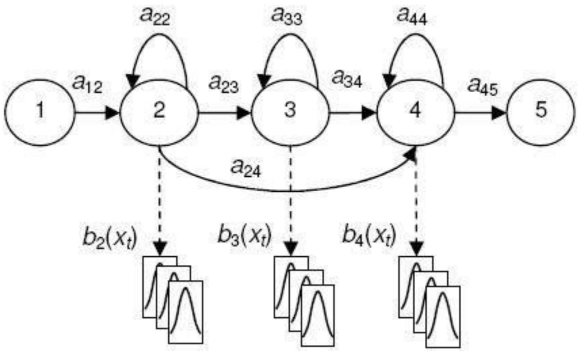
12

Your monthly repayment will be 130.48 RM

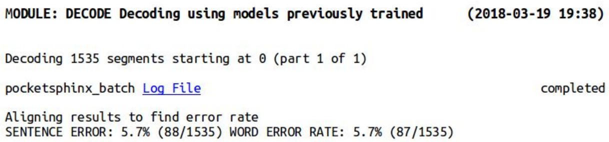
S

Do you want to continue?

Automatic Speech Recognition



https://www.gta.ufrj.br/grad/09_1/versao-final/impvocal/hmms.html

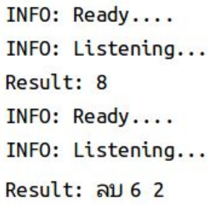


```
MODULE: DECODE Decoding using models previously trained      (2018-03-19 19:38)

Decoding 1535 segments starting at 0 (part 1 of 1)

pocketsphinx_batch Log File                                          completed

Aligning results to find error rate
SENTENCE ERROR: 5.7% (88/1535) WORD ERROR RATE: 5.7% (87/1535)
```



```
INFO: Ready....
INFO: Listening...
Result: 8
INFO: Ready....
INFO: Listening...
Result: ลบ 6 2
```

SERTIS

CONNECTING DATA TO ACTION

Jussi Jousimo
Natsuda Laokulrat
Pornbhussorn Kanchanakanok

jjousi@sertiscorp.com
www.sertiscorp.com