

wbs

WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

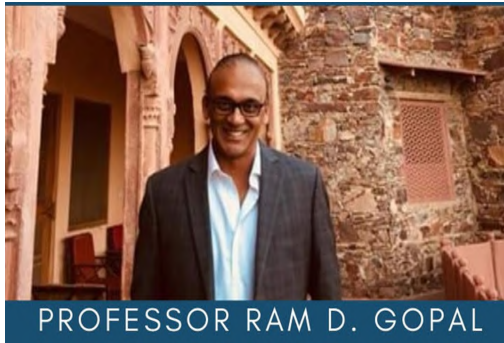
For the Open Minded

Ram Gopal
Michael Mortenson
Nikolaos Korfiatis (Marking)

2023-2024

Module Overview

IB9CW0
Text Analytics



Ram Gopal
Professor,
Information Systems Management and Analytics
Academic Director,
Gillmore Centre for Financial Technology
Warwick Business School

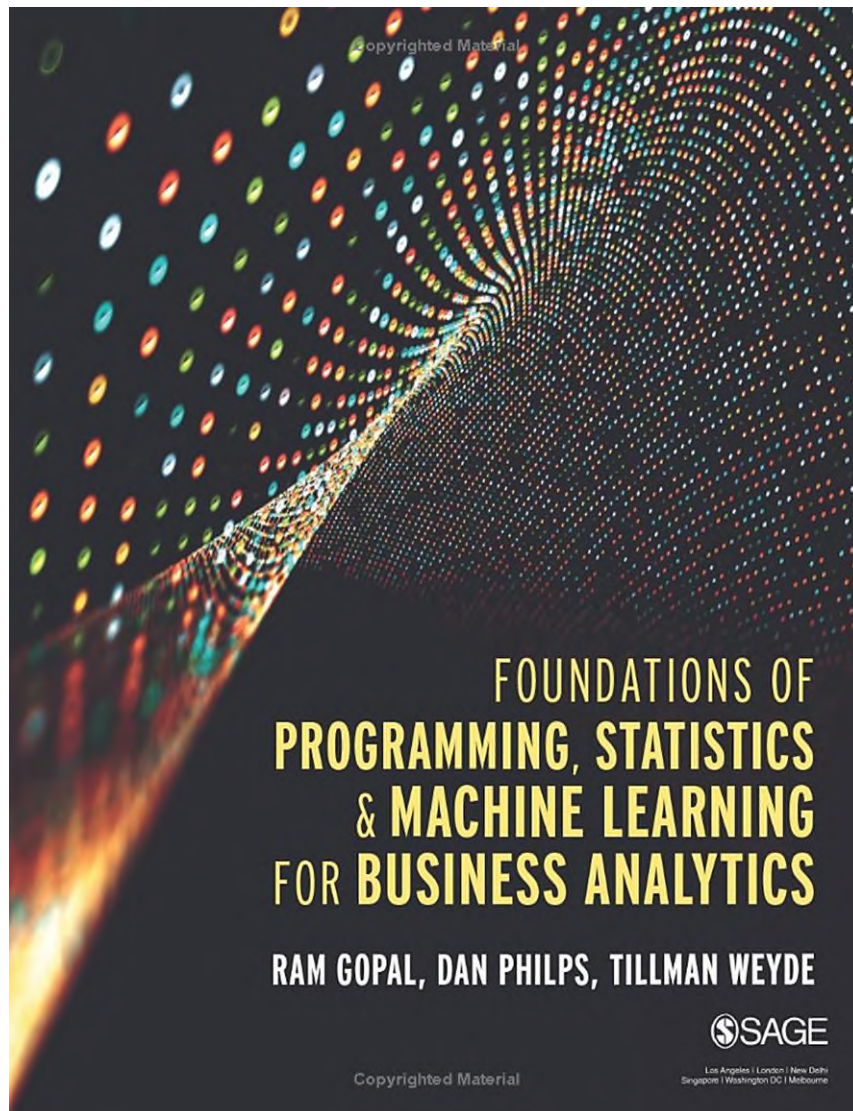
Email: ram.gopal@wbs.ac.uk

WBS Profile: <https://www.wbs.ac.uk/about/person/ram-gopal>

Google Scholar: <https://scholar.google.com/citations?user=6ytuTrkAAAAJ&hl=en&oi=ao>

Research Interests

Data Science; Predictive and Prescriptive Analytics; Financial Technologies; Health Informatics; Information Security, Privacy and Valuation; Economics of Intellectual Property Rights and Digital Goods; Online Market Design; Business Impacts of Technology



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

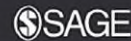
SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/1 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Ruth Stitt
Editorial assistant: Charlotte Hegley
Production editor: Ian Antcliff
Marketing manager: Lucia Sweet
Cover design: Naomi Robinson
Typeset by: C&M Digital (P) Ltd, Chennai, India
Printed in the UK



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Copyrighted Material

© Ram Gopal, Daniel Philps, & Tillman Weyde 2023

Apart from any fair dealing for the purposes of research, private study, or criticism or review, as permitted under the Copyright, Designs, and Patents Act, 1988, this publication may not be reproduced, stored, or transmitted in any form, or by any means, without the prior permission in writing of the publisher, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publisher.

Library of Congress Control Number: 2022947084

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-5296-2090-0
ISBN 978-1-5296-2091-7 (pbk)

At SAGE, we take sustainability seriously. Most of our products are printed in the UK using responsibly sourced papers and boards. When we print overseas, we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

Textbook GPT



Foundations Book Companion

Friendly academic instructor for 'Foundations of Programming, Statistics & Machine Learning for Business Analytics' by Ram Gopal, Dan Philps, and Tillman Weyde

Explain the concept of linear regression.

How is machine learning used in data analysis?

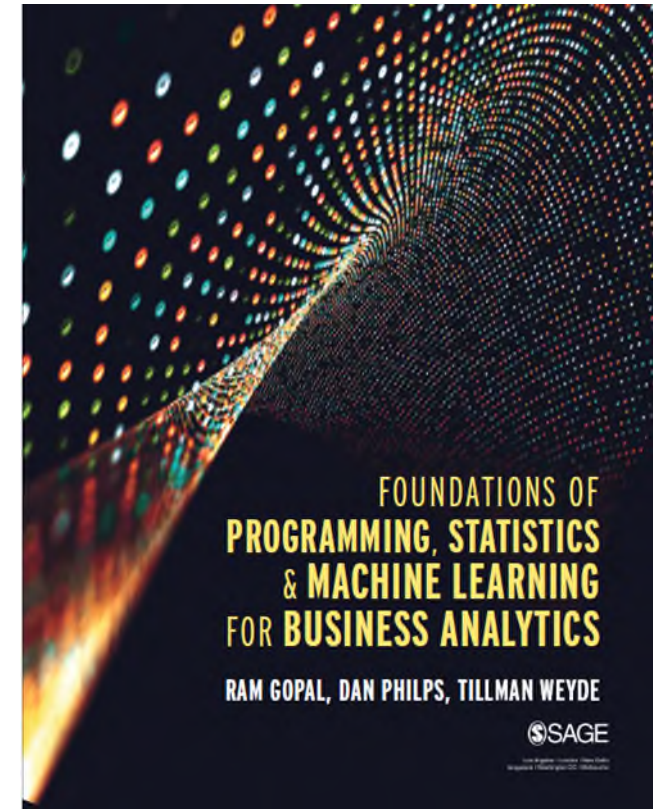
Show me a code snippet in Python for data visualizati...

Give me an R code example for statistical testing.

 Message Foundations Book Companion...



ChatGPT can make mistakes. Consider checking important information.



Textbook GPT

Foundations Book Companion ▾



You

Show me a code snippet in Python for data visualization.



Foundations Book Companion

Here's a Python code snippet for data visualization using the Seaborn library. This example demonstrates how to create a stacked histogram:

```
python Copy code

import seaborn as sns
import pandas as pd

# Assuming 'mpg' is a pandas DataFrame with columns 'class' and 'drv'
# You need to replace 'mpg' with your actual DataFrame name

# Sort values by 'class' and 'drv'
mpg_sorted = mpg.sort_values(by=['class', 'drv'])

# Create a stacked histogram
sns.histplot(data=mpg_sorted, x='class', hue='drv', multiple='stack', p
```

This code will produce a stacked histogram where the x-axis represents different car classes, and the colors (blue, red, green) represent different drive types (indicated by the 'drv' column). This kind of visualization is helpful for understanding the distribution of categories across different groups.



Message Foundations Book Companion...



ChatGPT can make mistakes. Consider checking important information.

LLM Oracle



LLM Oracle

LLM Guru: Explains concepts, creates diagrams, generates use cases, and offers Python snippets.

Explain the concept of tokenization in LLMs.

Provide a use case for GPT in healthcare.

Create a diagram showing the layers of a transformer...

Show me a Python snippet for text generation using ...

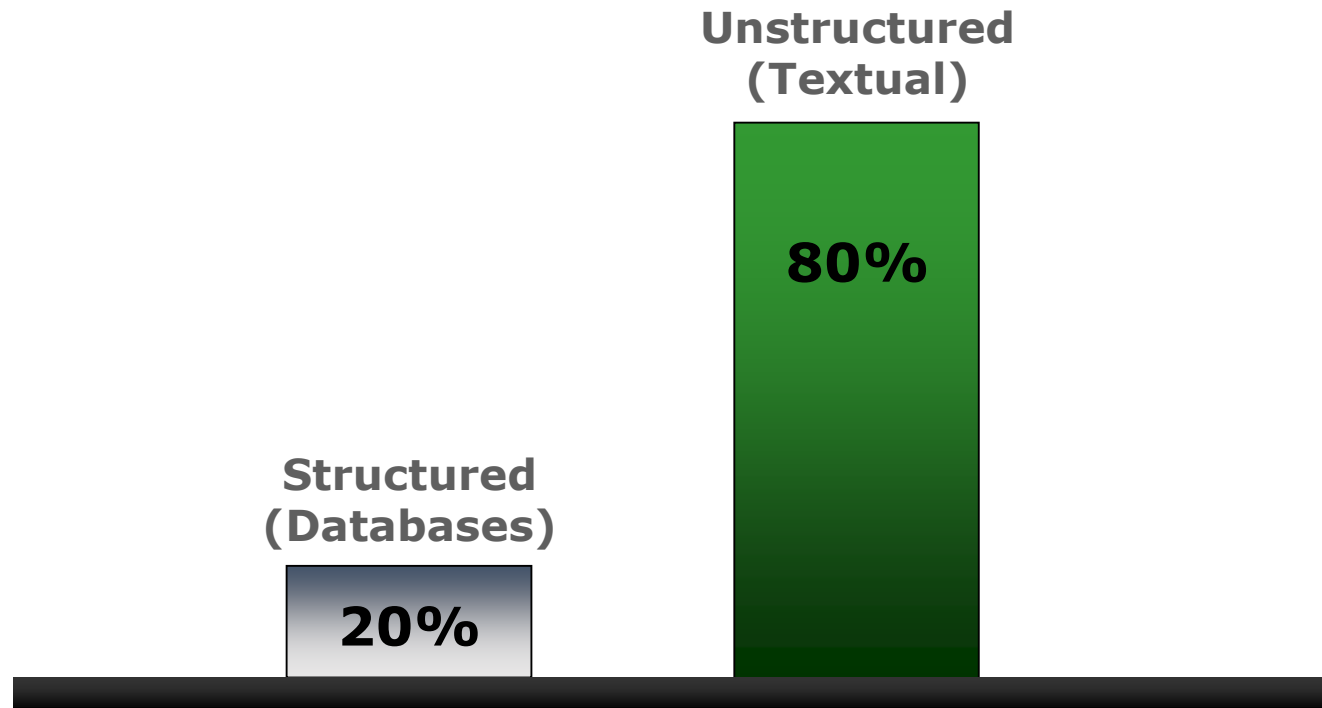
 Message LLM Oracle...



ChatGPT can make mistakes. Consider checking important information.

IB9CW0 Text Analytics

The Information Landscape



Natural Language Processing



SCENE FROM "DAN'L DRUCE."

This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant, is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Vezin.

- WHAT
- WHO
- ABOUT
- IMPLICIT

Natural Language Processing

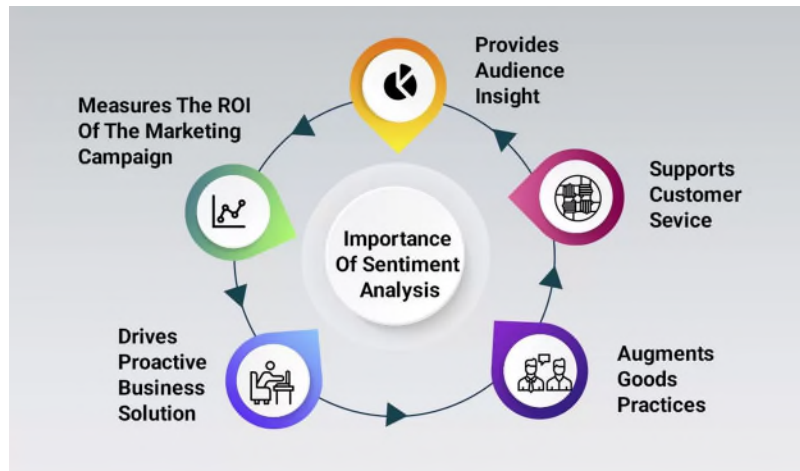
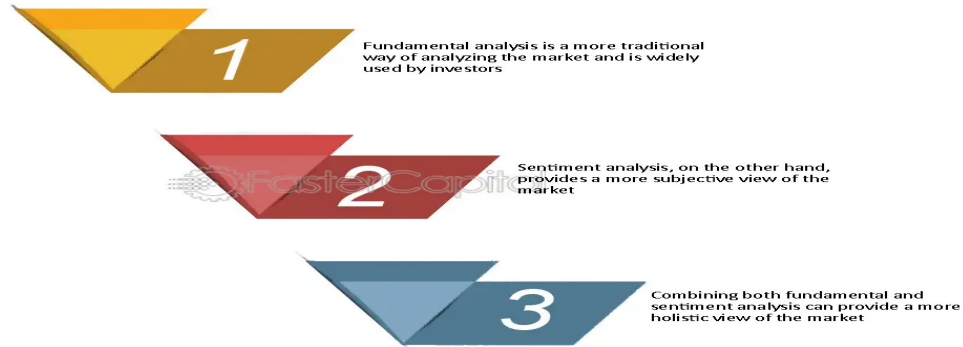


SCENE FROM "DAN'L DRUCE."

This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant, is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Vezin.

- WHAT → Sentiment Analysis, Opinion Mining
- WHO → Named Entity Recognition
- ABOUT → Feature Extraction
- IMPLICIT → Topic Modeling, Hidden Structures

Applications of Sentiment Analysis



The Importance of Sentiment Indicators in Trading



How Marketers Benefit from Sentiment Analysis Software



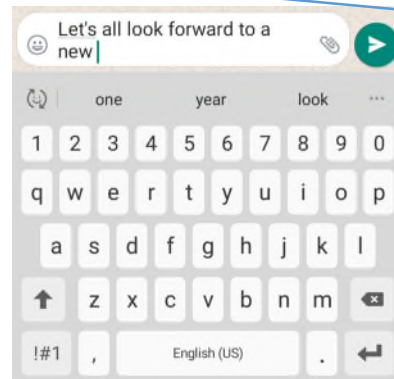
Language Models



SCENE FROM "DAN'L DRUCE."

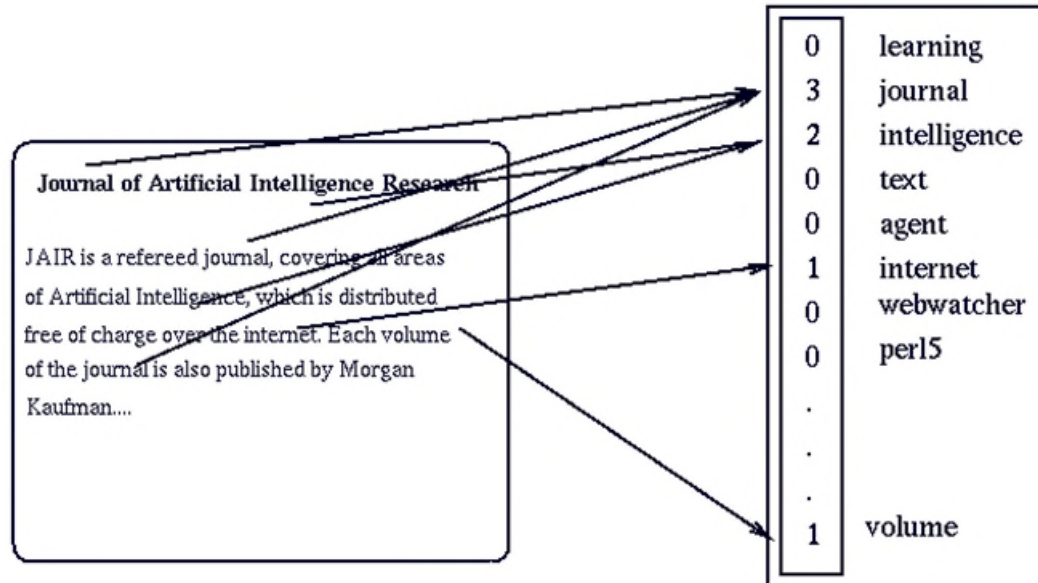
This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant, is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Vezin.

• WHAT'S NEXT?



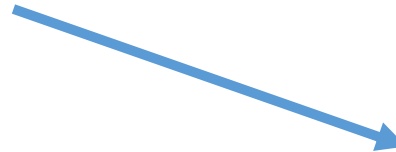
- ✓ Next word prediction
- ✓ Text generation
- ✓ Summarization

How do we work with text? Bag-of-words!



Tokenize

document	text
1	good product. Like it
2	product ready to use. use easy peasy peasy peasy
3	product of like work
4	not sure about product use



document	token
1	good
1	product
1	like
1	it
2	product
2	ready
2	to
2	use
2	use
2	easy
2	peasy
2	peasy
2	peasy
3	product
3	of
3	like
3	work
4	not
4	sure
4	about
4	product
4	use

Matrix Representation

document	text
1	good product. Like it
2	product ready to use. use easy peasy peasy peasy
3	product of like work
4	not sure about product use

document	peasy	use	good	it	like	product	easy	ready	to	of	work	about	not	sure
1			1	1	1	1								
2	3	2				1	1	1	1					
3					1	1				1	1			
4		1				1						1	1	1

We can now formally analyze text data – sentiment analysis, opinion mining, document similarity, clustering documents, etc.

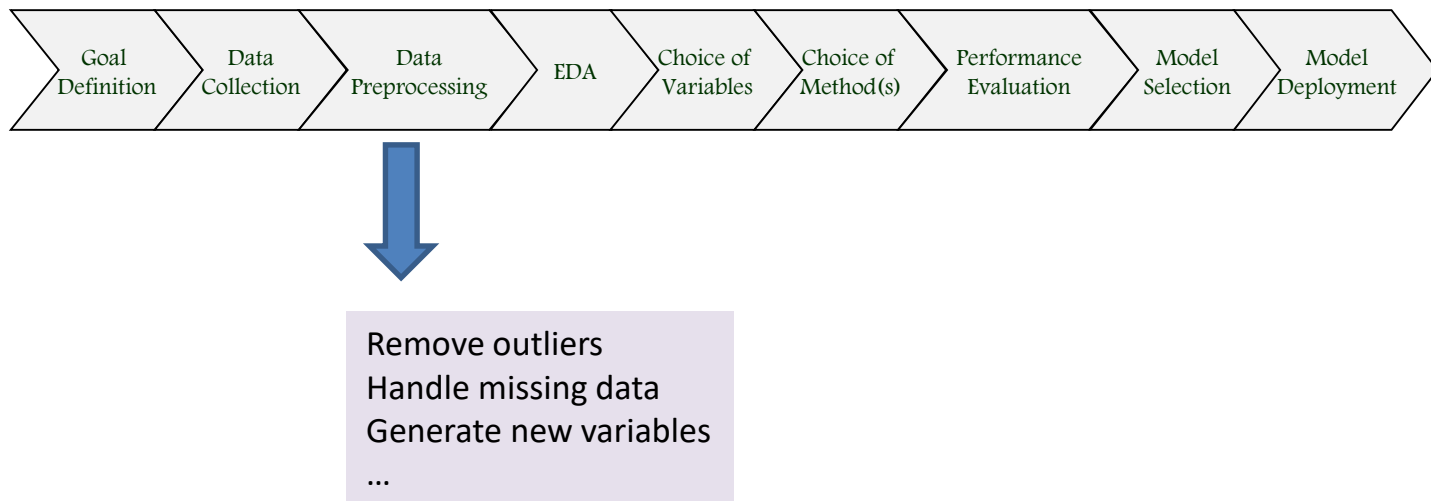
Data Mining (DM): transforms
structured data into actionable insights

Text Mining (TM): same, but
unstructured

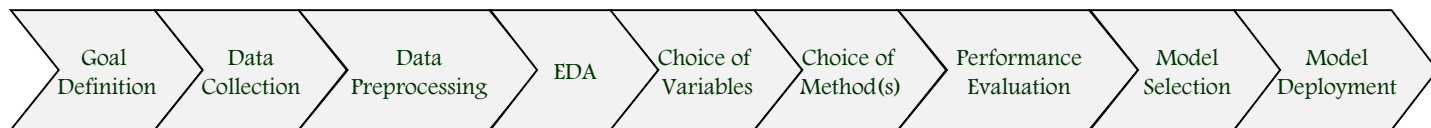
Text Mining (TM): transforms text into actionable insights

Natural Language Processing (NLP): same, but in a human-like fashion

The **data** mining process



The **text** mining process



Text into structure
Extract features from the text

Remove outliers
Handle missing data
Generate new variables
...

Text Mining: Examples

- Text mining is an exercise to gain knowledge from stores of language text.
- Text:
 - Web pages
 - Medical records
 - Customer surveys
 - Email filtering (spam)
 - DNA sequences
 - Incident reports
 - Drug interaction reports
 - News stories (e.g. predict stock movements)



Amazon Product Reviews

PLEASE DONT DO IT! this just rachett the palette size is not like in the picture, the colors go on sheer, slides on face like wax don't expect your makeup to stay put with this one. just spend the money for the good stuff. 2 thumbs down

The texture of this concealer pallet is fantastic, it has great coverage and a wide variety of uses, I guess it's meant for professional makeup artists and a lot of the colours are of no use to me but I use at least two of them on a regular basis, and two more occasionally, which is the only reason I'm giving it for stars, I feel like the range of colors is kind of a waste for me, but the product itself is wonderful, it's not cakey, gives me a natural for and concealed my imperfections, therefore I highly recommend it :)

I bought this for use on my face which has really bad Rosacea. It made things much worse and really did not help with moisture.

This palette was a decent price and I was looking for a few different shades. This palette conceals decently, however, it does somewhat cake up and crease.

Why dealing with Text is Easy

- Highly redundant data
 - ...most of the methods count on this property
- Just about any simple algorithm can get “good” results for simple tasks:
 - Pull out “important” phrases
 - Find “meaningfully” related words
 - Create some sort of summary from documents

Why dealing with Text is Hard – Ambiguity

Washington was born into slavery on the farm of Burroughs.

Washington went up 2 games to 1 in the four-games series.

Blair arrived in **Washington** for what may be his last state visit.

In June, **Washington** passed a primary seatbelt law.

The **Washington** had proved to be a leaky ship.

LLM sometime also cannot deal with all nuances of language. Eg in this case the Washington has different meaning in different sentences.

Why dealing with Text is Hard – Ambiguity

- Speech Recognition
 - “recognize speech” vs. “wreck a nice beach”
 - “youth in Asia” vs. “euthanasia”
- Syntactic Analysis words put into sequence
 - “I ate spaghetti **with** chopsticks” vs. “I ate spaghetti **with** meatballs.”
- Semantic Analysis meaning of the words. LLM does a good job
 - “The dog is in the **pen**.” vs. “The ink is in the **pen**.”
 - “I put the **plant** in the window” vs. “Ford put the **plant** in Mexico”
- Pragmatic Analysis
 - From “The Pink Panther Strikes Again”:
 - Clouseau: Does your dog bite?
Hotel Clerk: No.
Clouseau: [*bowing down to pet the dog*] Nice doggie.
[*Dog barks and bites Clouseau in the hand*]
Clouseau: I thought you said your dog did not bite!
Hotel Clerk: That is not my dog.

this example: humor => how machine can read that from the text

Text Mining Challenge – Ambiguity (cont'd)

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in n prepositional phrases has over 2^n syntactic interpretations.
 - I see the man *with* the telescope.
 - I see the man with the telescope. Dont know what is the true meaning thus we need the context
 - I see the man with the telescope.
- For some languages, even segmenting words is challenging!
 - 章魚的日文怎麼說?

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
 - One morning I shot an elephant in my pajamas. How he got into my pajamas, I'll never know.
 - She criticized my apartment, so I knocked her flat.
 - Policeman to little boy: "We are looking for a thief with a bicycle." Little boy: "Wouldn't you be better using your eyes."

Other challenges

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

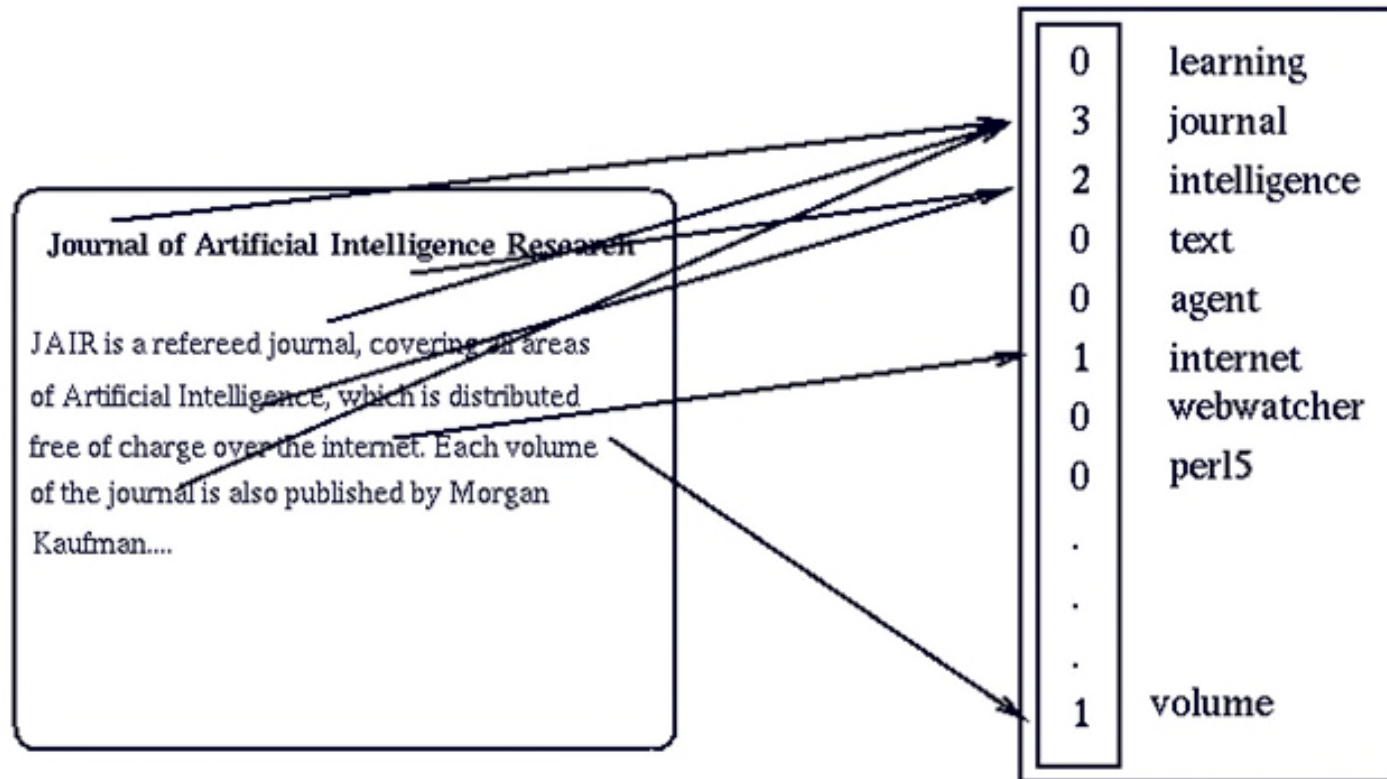
tricky entity names

Where is *A Bug's Life* playing ...
Let It Be was recorded ...
... a mutation on the *for* gene ...

Syntax, Semantic, Pragmatics

- Syntax concerns the proper **ordering** of words and its affect on meaning.
 - The dog bit the boy.
 - The boy bit the dog.
 - Bit boy dog the the.
- Semantics concerns the (literal) **meaning** of words, phrases, and sentences.
 - “plant” as a photosynthetic organism
 - “plant” as a manufacturing facility
 - “plant” as the act of sowing
- Pragmatics concerns the overall communicative and social **context** and its effect on **interpretation**.
 - The ham sandwich at the corner wants another beer. (co-reference, anaphora)
 - John thinks vanilla. (ellipsis)

Bag-of-words document representation



Sentiment Analysis



15 of 16 people found the following review helpful:

★★★★★ **Excellent intro to NLP**, July 17, 2009

By **P. H. Adams "phadams"** (Chesapeake, VA USA) - [See all my reviews](#)

REAL NAME

This review is from: *Natural Language Processing with Python (Paperback)*

Excellent introduction to the field of Natural Language Processing. I've been using the Natural Language Toolkit, the Python library explained in this book, for about two years and have seen it continually improve and become more robust. I eagerly awaited this text, which I first learned about over a year ago, and I must say the wait was worth it. Although most useful for those with a background in computer science or linguistics, it's a fairly gentle introduction to the field, so anyone with interest in the subject should find it useful and easy to understand. Stephen, Ewan, and Edward have done an excellent job of explaining language technologies and associated algorithmic functions for analyzing text.

Help other customers find the most helpful reviews

[Report abuse](#) | [Permalink](#)

Was this review helpful to you?

★★★★★ **It was our third time at Firefly**

 OpenTable Diner Since 2008

Dined on 09/18/2011

It was our third time at Firefly and once again it was an incredibly memorable meal. The food preparation was imaginative and the quality of the food was outstanding. The desserts were over the top.

Food	★★★★★
Service	★★★★☆
Ambiance	★★★★☆
Noise Level	Moderate

Special Features:

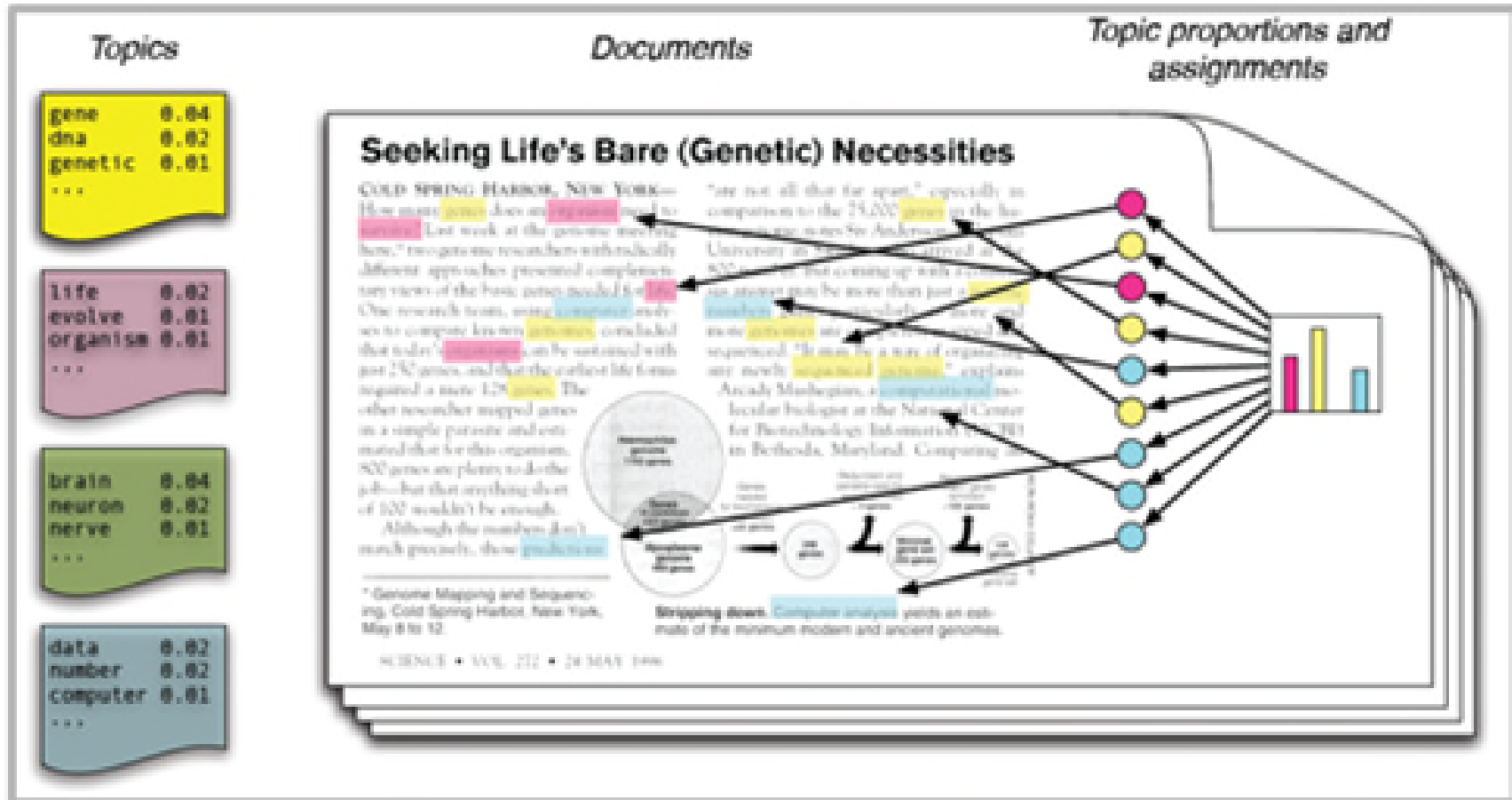
fit for foodies, neighborhood gem, notable wine list, special occasion

 [SHARE](#)

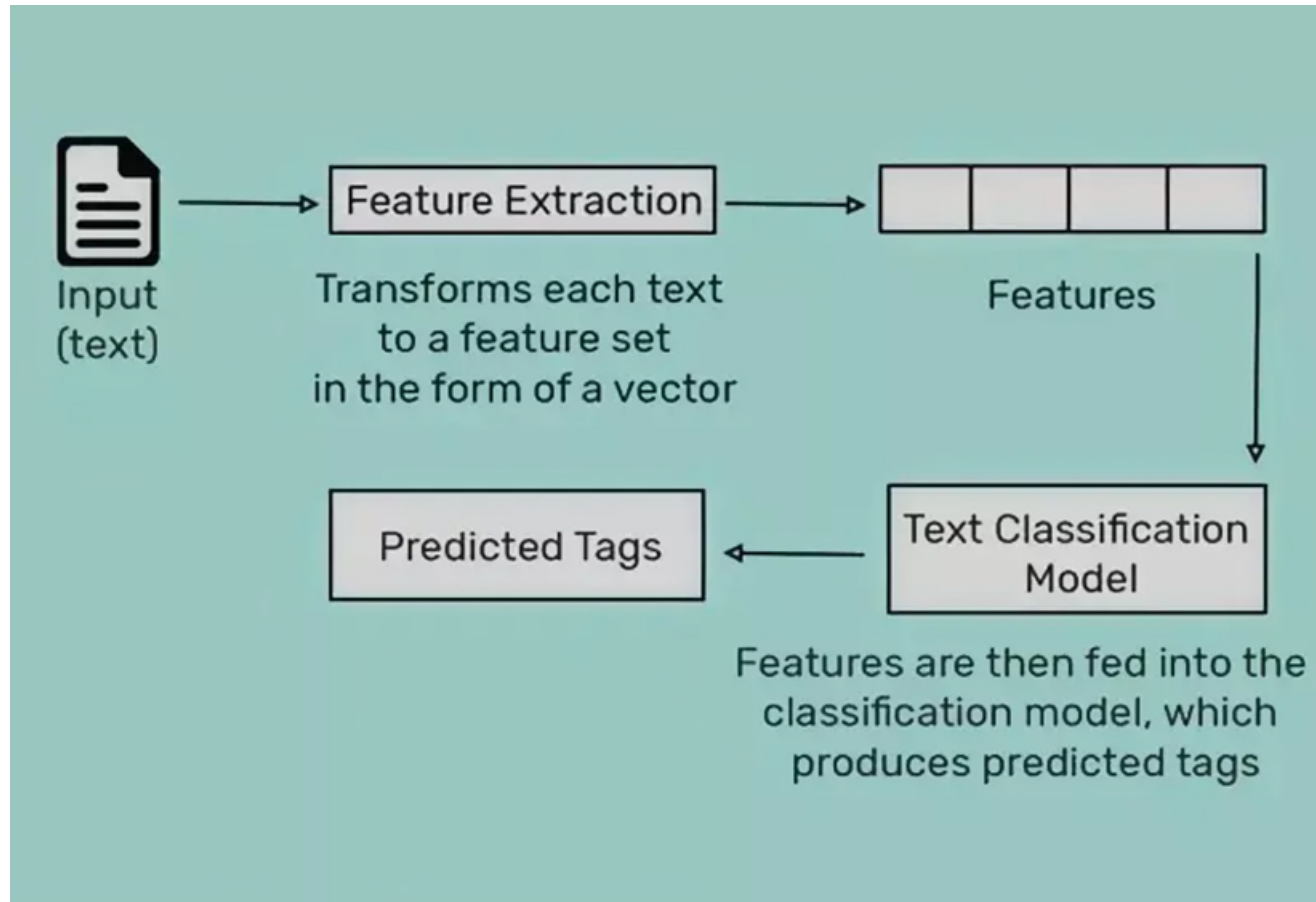
[Report inappropriate content](#)

- How opinions, attitudes, emotions, and perspectives are expressed in language provides a rich set of tools and techniques for extracting this evaluative, subjective information from large datasets and summarizing it.

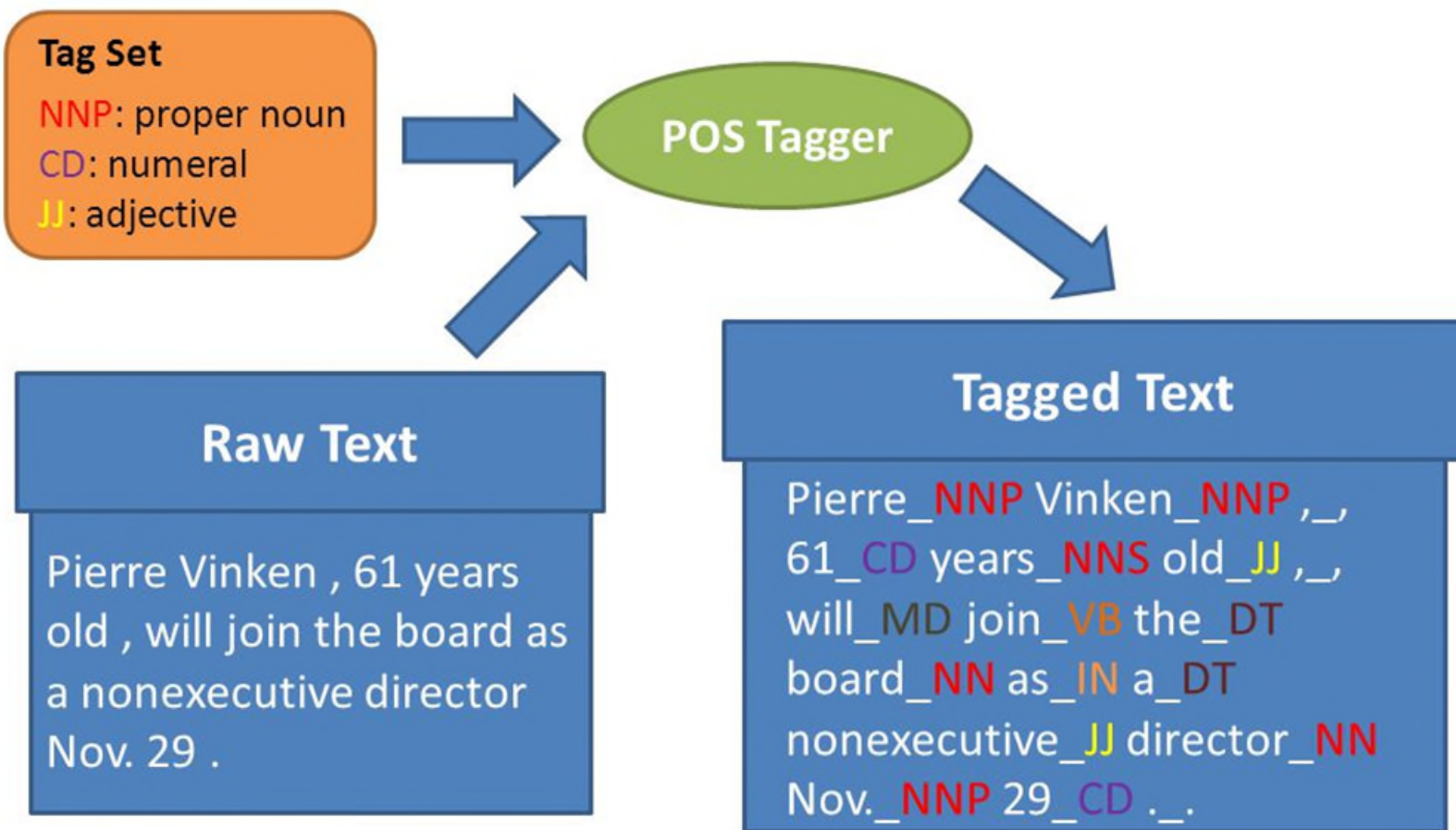
Topic Modeling



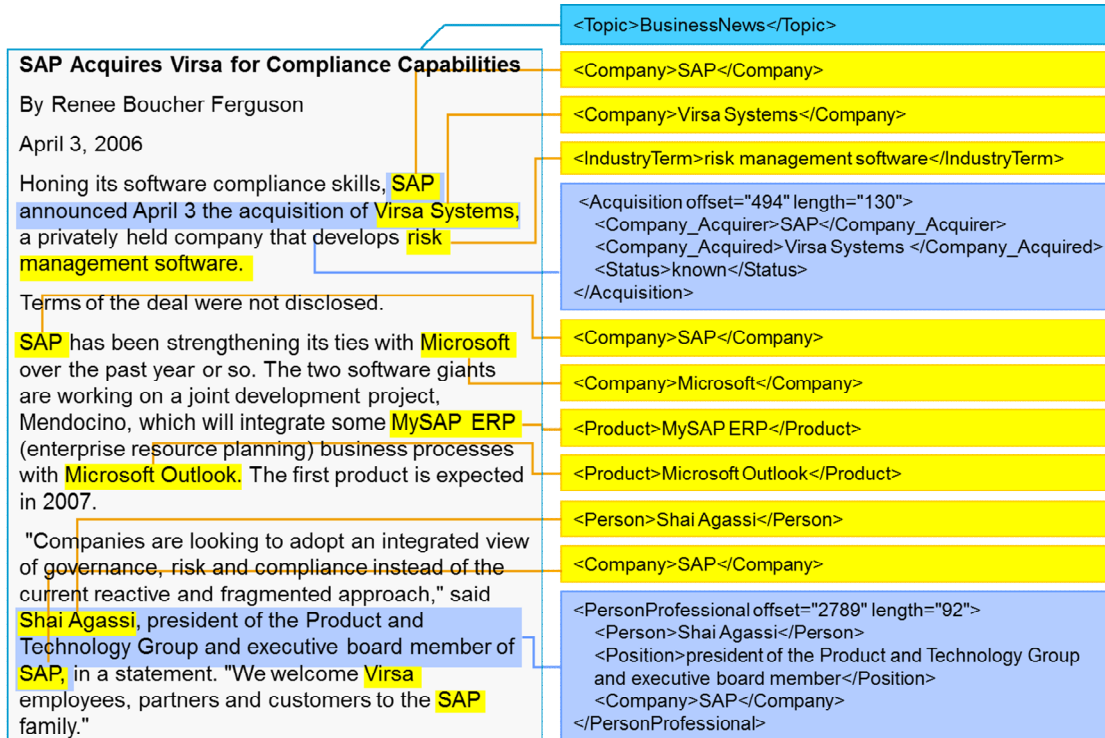
Supervised Learning



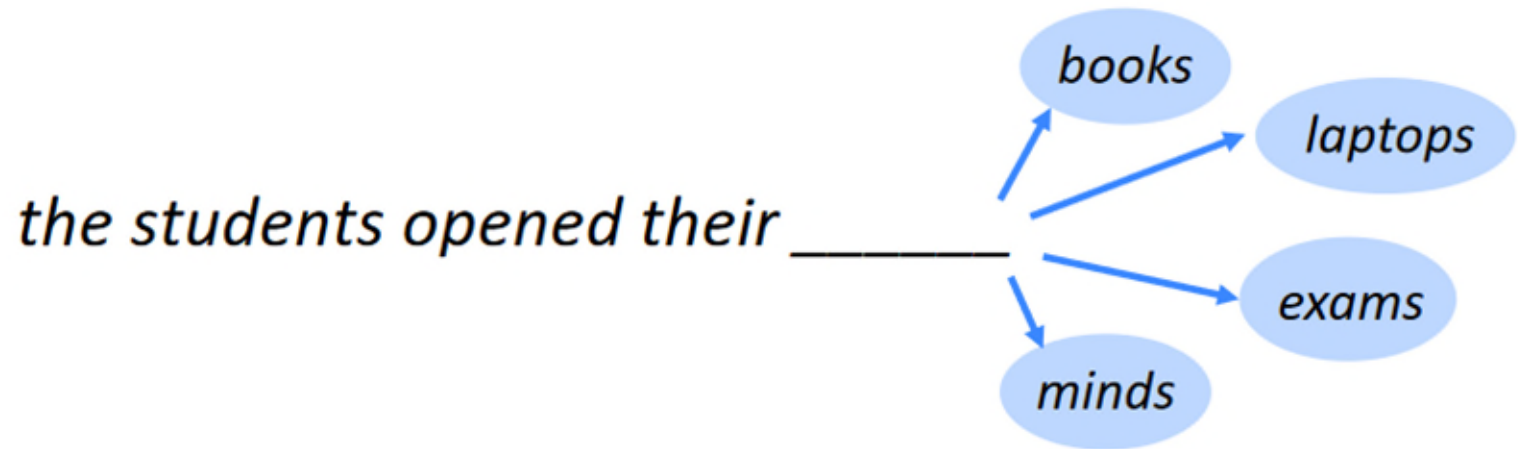
Parts of Speech (POS) Tagging



Named Entity Extraction



Language Models



Tentative Schedule of Topics

Topic	Details
Introduction to Text Mining	- regular expressions; text search and extraction
Tokenization and Pre-processing	- word, character, sentence, paragraph tokens; stop words, stemming, and lemmatization; n-grams; explore Google n-grams for word usage across time
Text Data Structures	- corpus and documents; word clouds, zip's law; TF, IDF, TF-IDF, DTM; word embeddings
Sentiment Analysis	- lexicons and lexicon-based sentiment scoring; valence shifter and language evolution (slang, emoticons, etc.)
Topic Modeling	- modeling with Latent Dirichlet Allocation (LDA) and embeddings
Supervised Learning with Text Data	- with term frequencies, topic models, and embeddings
Semantic Search with Embeddings	- computing sentence and paragraph embeddings from word embeddings; implementing semantic document search for natural language queries.
Language Models	- probabilistic models based on n-grams; next word prediction and generating text
Large Language Models	- fundamentals of LLMs; applications and inference with LLMs
Introduction to RAG (Retrieval Augmented Generation) systems	- key concepts; implementing a basic pipeline
RAG system development	-implementing end-to-end RAG pipelines; selecting embedding algorithms and chunking strategies; vector databases; choice of LLMs; query strategies; testing strategies; maintenance.
Advanced topics in LLMs	- multi-modal RAG/generation; AutoGPT/BabyAGI; agents; pathway to AGI.

Class Content and Structure

- Overview of theory and concepts
- *Learning-by-coding* approach. **Bring your laptops to class**
- Python in Colab environment
- Will use AI tools for coding assistance (ChatGPT, Claude, Gemini, Co-pilot, ...)
- GitHub (<https://github.com/RDGopal/IB9CW0-Text-Analytics>)

Assessment

- **Group assignment (40%)**
- **Individual assignment (60%)**

You will need to produce 2 outputs, 1500 words in total.

1. 1200-word essay
2. 300-word one page summary

Due Monday, 13 May 2024, before 12:00:00 (UK time)

1200-word essay:

The broad objective of the project is for student teams to either (a) explore advanced functionality and features on text analytics or (b) conduct an in-depth case study involving text analytics. Each group can discuss with the instructor regarding the appropriateness of the topic chosen.

The following short list provides some topics the groups may consider:

- Advanced language models
- text-to-**x** (where **x** could be audio, image, animation, video, etc.)
- **x**-to-text
- Text summarization
- Language translation
- Bayesian analysis in natural language processing
- Transformers
- Text sourcing
- Advanced document analysis
- Entity linking
- Prompt engineering

Marking Criteria

Completeness of the solution, efficiency of the code, interpretation of the results, and the quality of the narrative.

300-word one page summary

Each Group should submit a one-page executive summary of the project.

You will need to produce 2 outputs, 2000 words in total

1. 1500-word essay
 2. 500-word commentary on a randomly assigned group's project (other than yours)
-

1500-word essay: **Due Thursday, June 6, 2024**

The goal of this assignment is to build a retrieval augmented generation (RAG) system, based on a large language model (LLM). The RAG system should be based on a particular domain of your choice, and your system should be able to combine both the generative/informational power of large language models, along with an ability to retrieve specialised information that would not be in the training data of the LLM (i.e. it should be able to generate content based on knowledge that was not publicly available at the point of training).

Broadly speaking, your approach should incorporate the following steps:

1. Identify a relevant domain in which to work. This can range from a more general area, such as academic research in a specific topic, through to something more company specific, e.g. using company-specific data/information.
2. Source a dataset/knowledge-base to use in your RAG system. This should be a set of documents, and you can use a publicly available dataset (e.g. from Kaggle).
3. Determine a set of tests (e.g. queries) that the RAG system should be able to achieve if it is to perform as a working system (and out perform the LLM alone).
4. You will need to chunk and embed the knowledge-base (using a pretrained embedding model of your choice) and store it in a vector database.
5. Develop an appropriate prompt template to query the LLM.
6. Create a query pipeline that will process (embed) a query; retrieve relevant information from the vector database; present the query and additional information to the LLM as a prompt template; and return the generated content.
7. Test the system thoroughly and document any future improvements you would make.

500-word commentary: Due Tuesday, 21 May 2024

You will be randomly assigned to one project done by other groups. You need to answer the following questions:

1. Describe the main strengths and weaknesses of the writeup. Is it easy to follow? Are the examples and illustrations appropriate? Is something important missing?
2. Was the project substantially original? Did the team appear to 'copy and paste' code and the substantial parts of the report from online sources? What sources did you check?
3. Is this better than your project? Why or why not?
4. Your rating (scale of 1 to 10)