

COMP5046

Natural Language Processing

Lecture 1: Course Info & Introduction to NLP

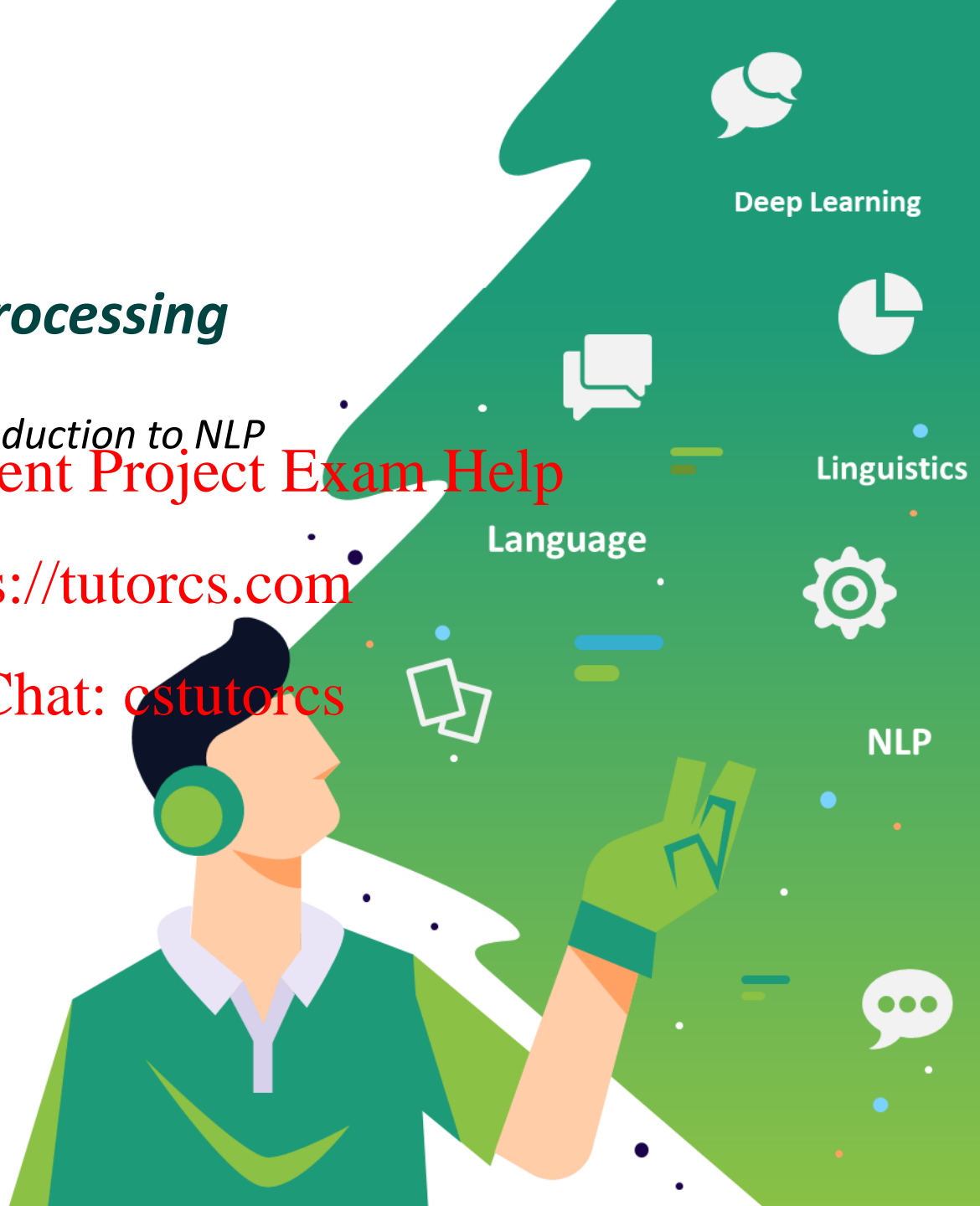
Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Dr. Caren Han

Semester 1, 2021
School of Computer Science,
University of Sydney



Lecture 1: Introduction to Natural Language Processing

1. *Course Introduction*

2. *Overview of Natural Language Processing (NLP)*

3. *Word Meaning and Representation*

4. *Count-based Word Representation*

- *One-hot Encoding*
- *Bag of Words*
- *Term Frequency-Inverse Document Frequency*

5. *Next Week Preview*

- *Prediction-based Word Representation*

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Dr Caren Han

Education

- B CompScience (1st Class Honours)
- PhD Computer Science (Artificial Intelligence)

Assignment Project Exam Help

Australian Young Achiever Teaching Excellence Award

Teacher of the Year 2020 Award

- Machine Learning, Natural Language Processing, etc.

<https://tutorcs.com>
WeChat: cstutorcs

NLP Research Experience (8 years experience)

- **U.S. Air Force:** Robotics Natural Language Processing Interface
- **U.S. Navy:** Deep Learning with NLP
- **Hyundai Co.:** Car Navigation Dialog System
- NLP Top-tier Conferences Best Paper Award/Best Area Paper Award

COMP5046 Natural Language Processing

This unit introduces computational linguistics and the statistical techniques and algorithms used to automatically process natural languages. It will review the core statistics and information theory, and the basic linguistics, required to understand natural language processing (NLP).

NLP is used in a wide range of applications, including information retrieval and extraction; question answering; machine translation; and classifying and clustering of documents. This unit will explore the key challenges of natural language to computational modelling, and the state-of-the-art approaches to the key NLP sub-tasks, including tokenisation, morphological analysis, word sense representation, part-of-speech tagging, named entity recognition and other information extraction.

Students will implement many of these sub-tasks in labs and assignments, that can be used in the real-world cases. The unit will also investigate the annotation process that is central to creating training data for interesting application. With this unit, students can develop the innovative application that can be used in the real world.

Where to find the course information?

Unit Outline - COMP5046

<https://www.sydney.edu.au/units/COMP5046>

Canvas – COMP5046

<https://canvas.sydney.edu.au/courses/30912>

What will you learn in this course?

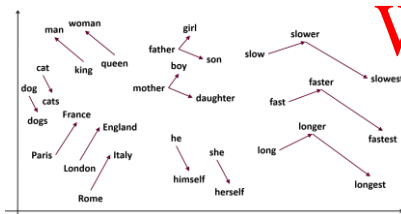
The focus of this course is on the review and comparison of models and methods that have achieved state-of-the-art results on various NLP tasks such as question answering (QA) and machine translation.

In this comprehensive review, students will get a detailed understanding of the past, present, and future of NLP. In addition, students will learn some of the current best practices for applying deep learning in NLP

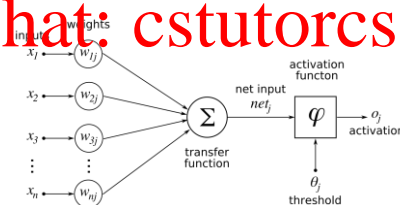
Assignment Project Exam Help

<https://tutorcs.com>

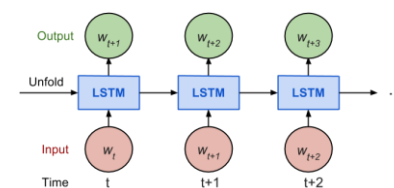
WeChat: cstutorcs



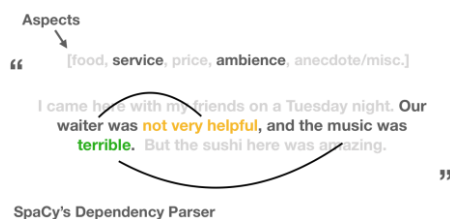
word2vec



NN

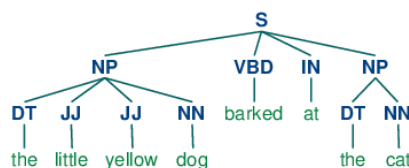


Language Modelling

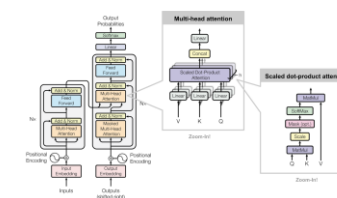


SpaCy's Dependency Parser

Dependency Parsing



Part-of-Speech Tagging



Transformer

INTRODUCTION

What will you learn in this course?

Week 1: Introduction to Natural Language Processing (NLP)

Week 2: Word Embeddings (Word Vector for Meaning)

Week 3: Word Classification with Machine Learning I

Week 4: Word Classification with Machine Learning II

*NLP and
Machine
Learning*

Week 5: Language Fundamentals

Week 6: Part of Speech Tagging

Week 7: Dependency Parsing

Week 8: Language Model and Natural Language Generation

*NLP
Techniques*

Week 9: Information Extraction: Named Entity Recognition

Week 10: Advanced NLP: Attention and Reading Comprehension

Week 11: Advanced NLP: Transformer and Machine Translation

Week 12: Advanced NLP: Pretrained Model in NLP

*Advanced
Topic*

Week 13: Future of NLP and Exam Review

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

I DO assume you can program

- By that, I mean you are a confident programmer
- Labs will involve programming
- Assessment will involve programming
- Python recommended; other popular languages accepted
- There will be **NO NON-programming option** for assignments
- But it's more than just programming:
 - algorithms, mathematics and (esp.) statistics
 - linguistics and intuition about language
 - analytical thinking

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



I DO NOT assume you are a linguist

- But you do **need to know** roughly **how to identify a noun/verb/etc.**
- We will think critically about **how we use language**
- and about how computational models capture **aspects of language**

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



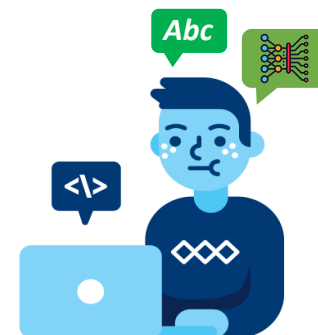
I DO NOT assume you are a deep learning researcher

- But you do need to know (really) **roughly how machine learning works.**
- We will think critically **how to use text data and embeddings**
- and about how deep learning models capture **aspects of language (context)**

Assignment Project Exam Help

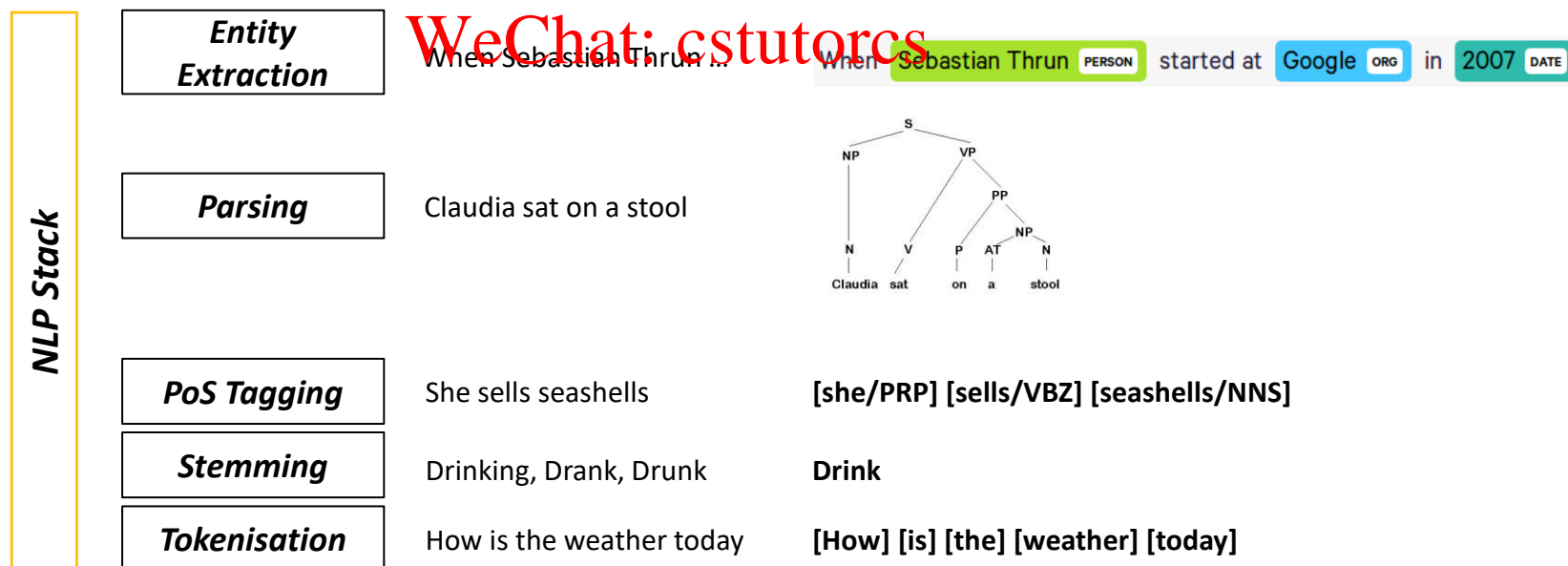
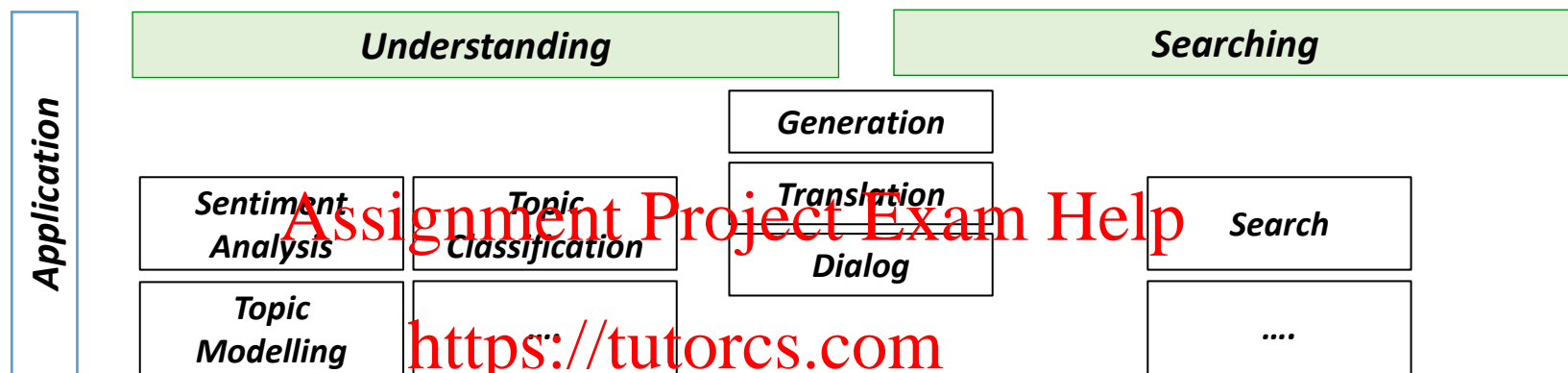
<https://tutorcs.com>

WeChat: cstutorcs

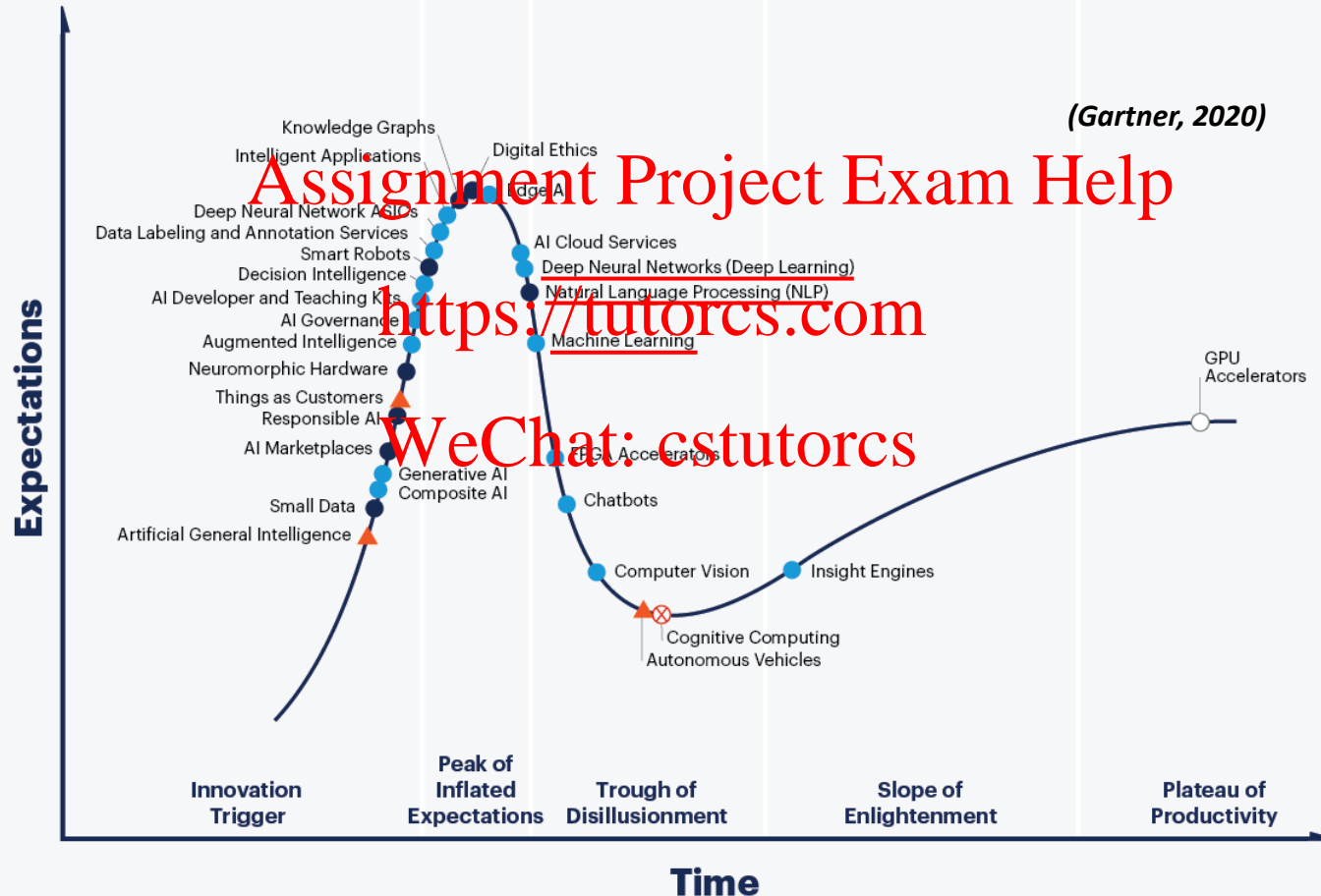


1 The NLP Big Picture

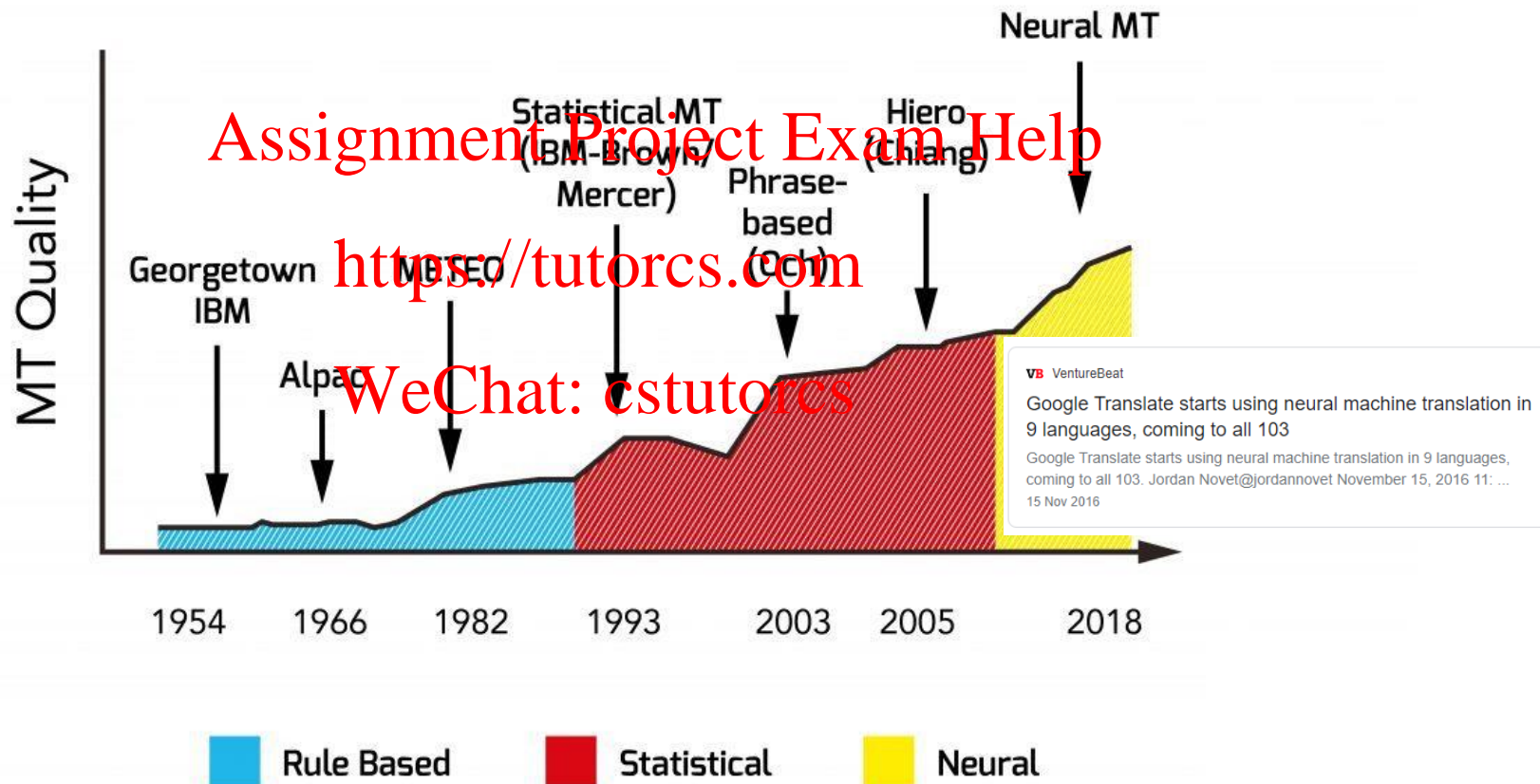
The purpose of Natural Language Processing: Overview



Hype Cycle for Artificial Intelligence

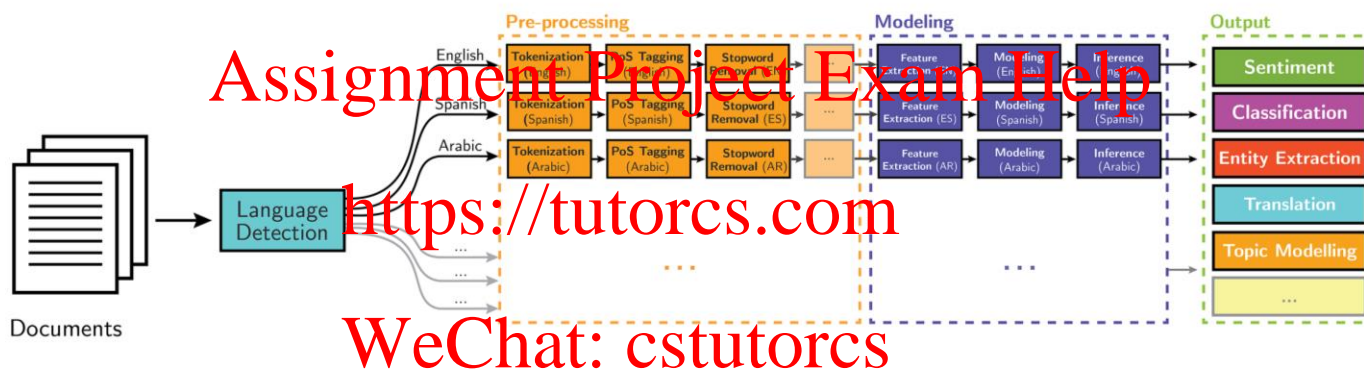


NLP Techniques – with the Trend of Machine Translation

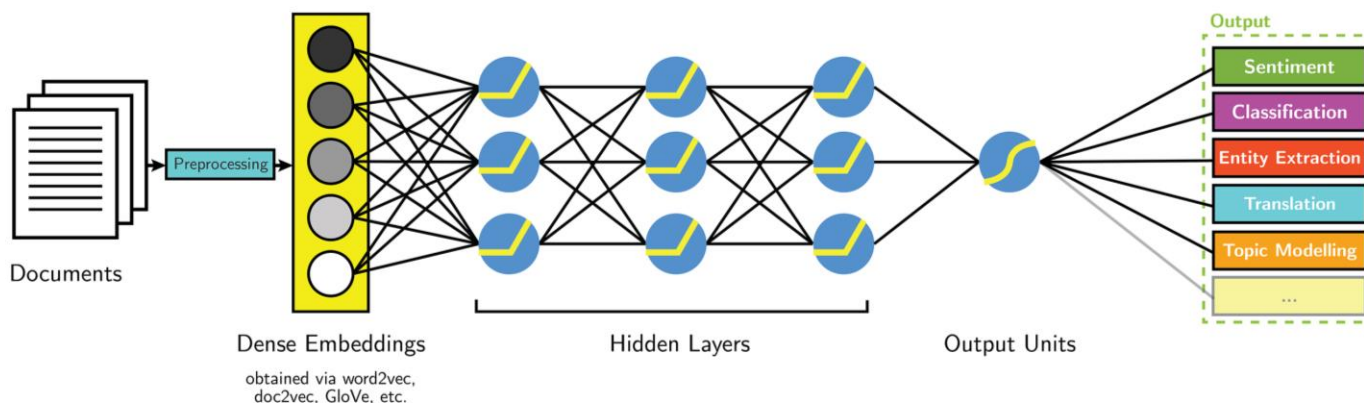


Language Modelling using Deep Learning-based NLP Techniques

Classical NLP



Deep Learning-based NLP



Assessment Overview

| Assessment | Weight | Due |
|--------------|--------|------------------------------------|
| Lab Exercise | 10% | Multiple Weeks |
| Assignment 1 | 20% | Week 8 (Friday 11.59pm – AU time) |
| Assignment 2 | 20% | Week 14 (Friday 11.59pm – AU time) |
| Final Exam | 50% | Exam Period |

<https://tutorcs.com>

Lab Exercises

- Programming tasks done in weekly computer labs

Assignments

- Take place through the teaching period
- Implementation and Documentation

Penalties for lateness

- 10% of the available marks per day late; maximum 7 days late

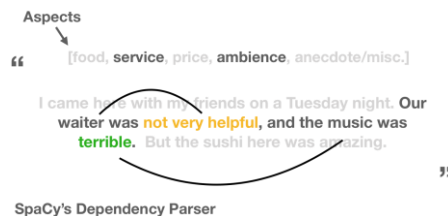
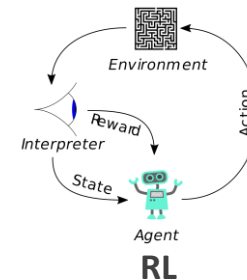
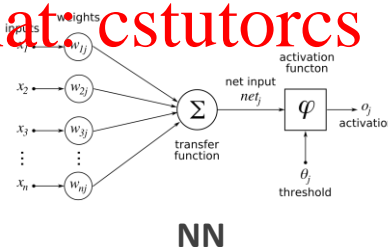
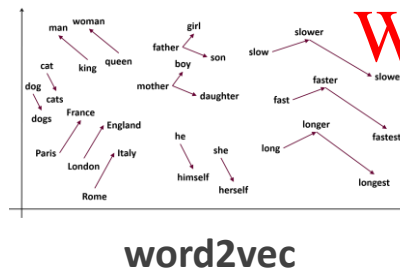
Lab Exercise – 10%

- In the Lab, students need to do the small tasks (1% for each week).
- 2-3 tasks are given based on what you learned in the previous lecture.
- You must have been assessed as having completed 10 out of 11 in order to get the 10% for weekly lab exercise.

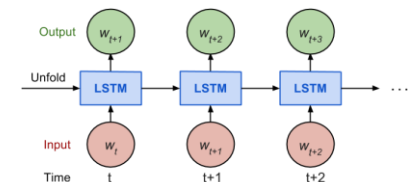
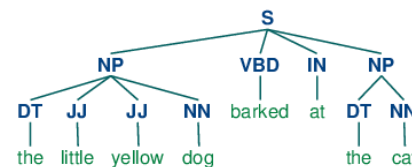
Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



Dependency Parsing

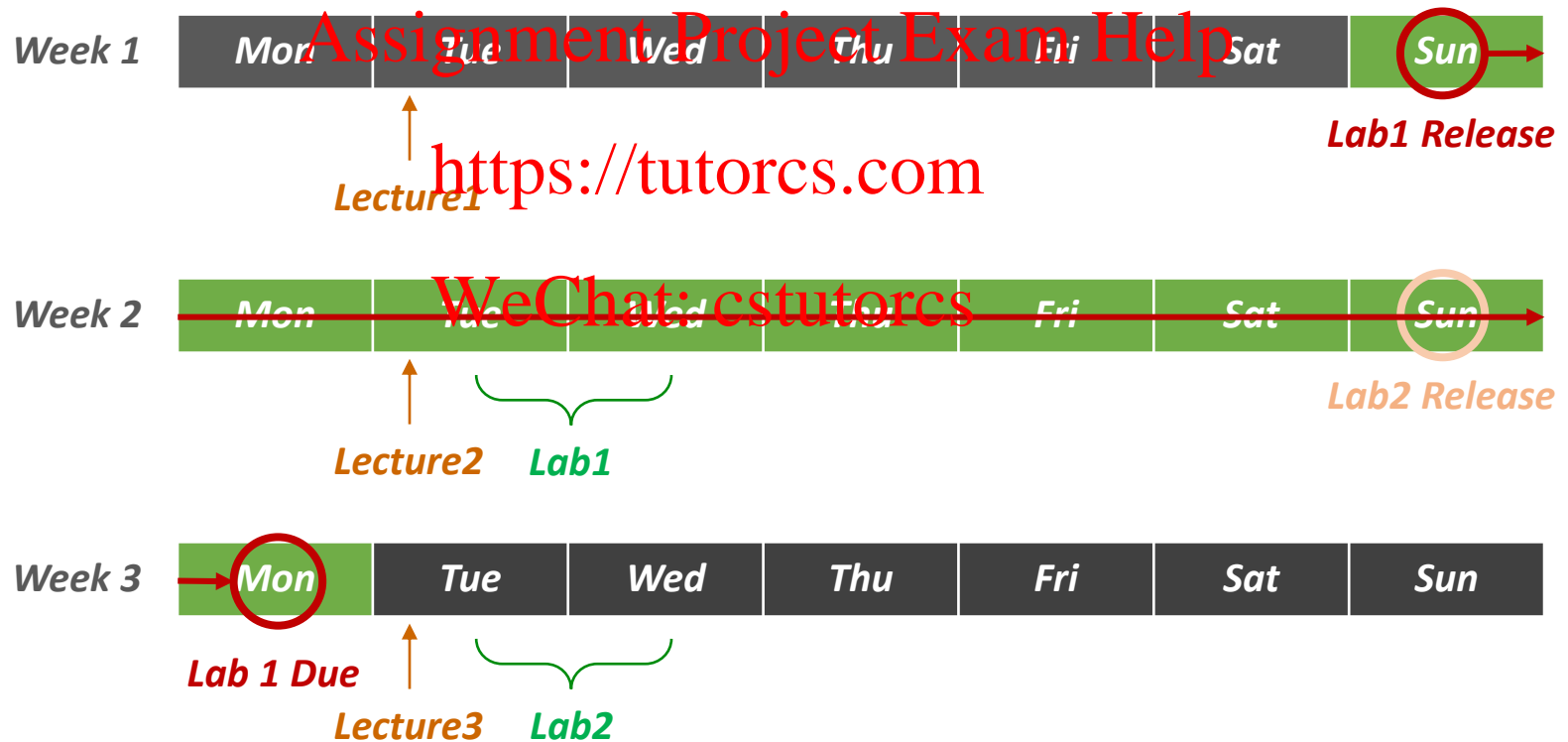


1

ASSESSMENT

Lab Exercise – 10%

- When to submit the Weekly Lab Exercise (e.g. Lab1 Release and Submission)



What do we do during Labs?

In Labs, Students will use Google CoLab

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.

Assignment Project Exam Help



Assignment 1 (20%)

The focus of this assignment 1 is implementing the text/word embeddings/language models, which **1)understands the language** and **2)produces the detection/prediction/generation** decision.

NOTE: Assignment 1 will be an individual assignment.

Assignment 2 (20%)

The focus of this assignment 2 is proposing a natural language model to produce high performance (such as accuracy, consistency, plausibility, validity, and distribution) in different NLP tasks. The different NLP tasks can be Question Answering, Named Entity Recognition, Text Classification, or etc.

NOTE: Assignment 2 would be a group assignment (2 people in group).
However, you can do individually only if you want.

Final Exam (50%)

The final exam will be a short take-home exam hosted on Canvas (3 hours duration). You will be asked to answer variety of theoretical questions.
The sample exam questions will be shared in the week 13 lecture.

ASSESSMENT Due

Week 1: Introduction to Natural Language Processing (NLP)

Week 2: Word Embeddings (Word Vector for Meaning)

Week 3: Word Classification with Machine Learning I

Week 4: Word Classification with Machine Learning II

Week 5: Language Fundamentals

Week 6: Part of Speech Tagging

Week 7: Dependency Parsing

Week 8: Language Model and Natural Language Generation

Week 9: Information Extraction: Named Entity Recognition

Week 10: Advanced NLP: Attention and Reading Comprehension

Week 11: Advanced NLP: Transformer and Machine Translation

Week 12: Advanced NLP: Pretrained Model in NLP

Week 13: Future of NLP and Exam Review

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Assignment 1 Due

Assignment 2 Due (Week 14)

Start assignments early!

- All assignments involve coding and report writing
- Reports are the primary deliverable
- Though we will check implementations for correctness
- Assignments will Be very different from last year's
- Reports will be submitted to Turnitin through Canvas
- Code is also submitted (for assignments 1 and 2) and retained
- We will use code plagiarism detection tools
- Clearly reference any copied/adapted code portions and cite their origins

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Start assignments early!

- Starting early means you will work while you sleep
- ~~Don't waste all your time on code~~
- The report is more important, but largely depends on the code
- If you're stuck, ask early
- We might be able to offer you alternatives

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Working Hours

- Work 12 hours per week for this course (including 3 contact hours);
- Attend 2 hours of lectures per week
 - Tuesday 3 – 5pm
 - Lectures are recorded, but don't depend on it!
- Attend 1 hour of tutorial/lab/prac time
- Participate respectfully in discussions in lectures and labs;
- Complete all assessment tasks on time.

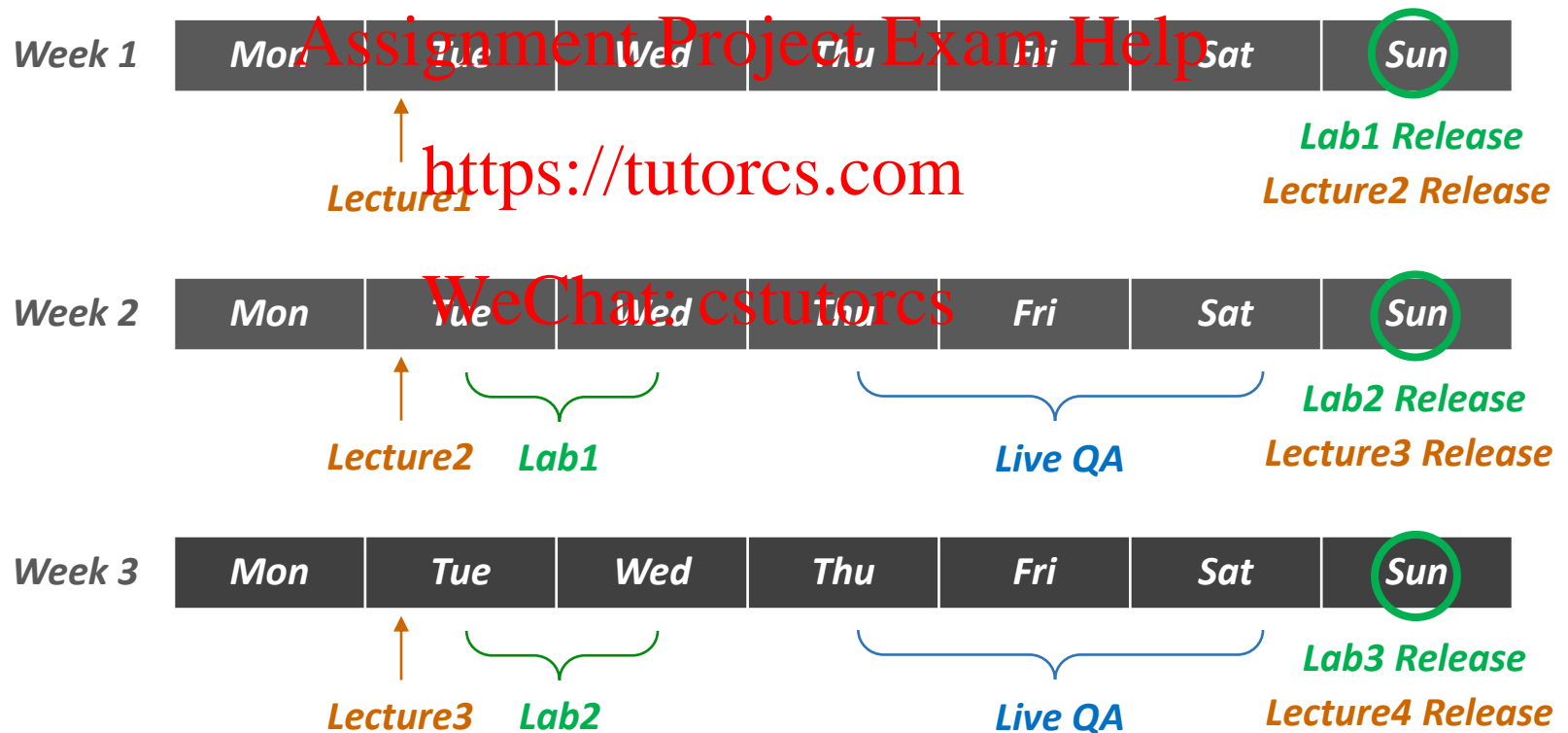
Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Classes and Release Date

- When the Class and Release Dates are



Full Course Timetable at course website:

- **Lecture:** Tue 3-5pm
- **Tutorial:** Tue 5-6pm, 6-7pm / Wed 5-6pm, 6-7pm
- **LiveQA:** Thu, Fri, Sat 9-10pm

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Who are we?

Unit Coordinator & Lecturer

- Caren.Han@sydney.edu.au
- No fixed consultation hour, please arrange a time to see me.

Assignment Project Exam Help

<https://tutorcs.com>

Teaching Assistants & Tutors

WeChat: cstutorcs

- They will introduce themselves really soon!

For Qs related to course content, please post in Ed.

For Qs related to admin, please contact to the unit coordinator

- Please put [COMP5046] in the title of the email

No Textbook Recommended, but if you really want to read some

Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.

Assignment Project Exam Help

Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. "O'Reilly Media, Inc.".

<https://tutorcs.com>

Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

WeChat: cstutorcs

Ask yourself...

- How much work will you be devoting to this unit, each week?
- Who should you see if difficulties arise?
- When is the first assessment due?
- What do you do if you get sick during semester?
- What is Turnitin?
- What programming language do you need to know?

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

What do we learn? (NLP and Deep Learning)



WeChat: cstutorcs

"We have an incompetent administration, and if I am not elected President, that will not change over the next four years—but it must change, and it must change now."

(June 13th, 2017)

We fed 270,000 words spoken by Trump into a computer program that studies language patterns.

"Now, they do charge you tariff on trucks, when we send trucks and other things over there."

(Announcement June 16th, 2017)

<https://www.youtube.com/watch?v=EFHyzuqjaok>

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. **Overview of Natural Language Processing (NLP)**
3. Word Meaning and Representation
4. Count-based Word Representation
 - One-hot Encoding
 - Bag of Words
 - Term Frequency-Inverse Document Frequency
5. Next Week Preview

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Why Process Language?

- language stores knowledge
- language communicates new knowledge
- language is a key to culture and human experience
- language is a natural interface for humans

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



Why Process Language?

- language stores knowledge
- language communicates new knowledge
- language is a key to culture and human experience
- language is a natural interface for humans

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



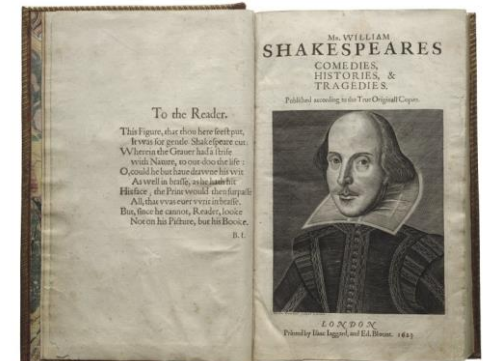
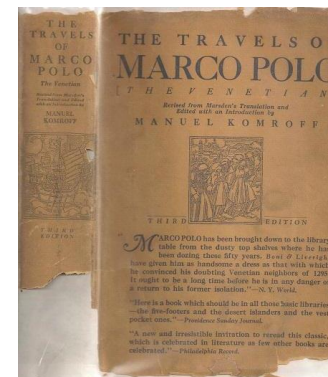
Why Process Language?

- language stores knowledge
- language communicates new knowledge
- language is a key to culture and human experience
- language is a natural interface for humans

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



Why Process Language?

- language stores knowledge
- language communicates new knowledge
- language is a key to culture and human experience
- **language is a natural interface for humans**

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



What is Natural Language Processing?

Natural Language Processing (NLP) is the branch of artificial intelligence focused on developing systems that allow computers to communicate with people using everyday language

Assignment Project Exam Help

Computational Linguistics

It concerns how computational methods can aid the understanding of language

WeChat: cstutorcs

Communication

The goal in the production / comprehension of language is communication.

Communication for the speaker:



Assignment Project Exam Help

<https://tutorcs.com>

- Intention: **Decide when and what information** should be transmitted (a.k.a. strategic generation). May require planning and reasoning about agents' goals and beliefs.
- Generation: **Translate the information to be communicated** (in internal logical representation or "language of thought") into string of words in desired natural language (a.k.a. tactical generation).
- Synthesis: **Output the string** in desired modality, text or speech.

WeChat: cstutorcs

Communication for the hearer:



Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

- Perception: **Map input modality to a string of words**, e.g. optical character recognition (OCR) or speech recognition.
- Analysis: **Determine the information content of the string.**
 - Syntactic interpretation (parsing): Find the correct parse tree showing the phrase structure of the string.
 - Semantic Interpretation: Extract the (literal) meaning of the string .
 - Pragmatic Interpretation: Consider effect of the overall context on altering the literal meaning of a sentence.
- Incorporation: Decide whether or not to **believe the content of the string** and add it to the Knowledge Base.

What is special about NLP?

- Human language is a system specifically constructed to **convey meaning** and is **not produced by a physical manifestation of any kind**. In that way, it is very different from vision or any other machine learning task.

Assignment Project Exam Help

- Most words are just symbols for an extra-linguistic entity : the word is a signifier that maps to a signified (idea or thing).

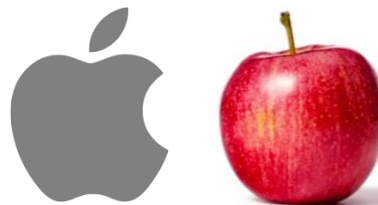
<https://tutores.com>

WeChat: cstutores

“Computer”



“Apple”



“Whaaaaaaa”

???????

Ambiguity

I saw the man on the hill with a telescope.

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs



Ambiguity is Explosive

Ambiguities compound to generate enormous numbers of possible interpretations.
In English, a sentence ending in n prepositional phrases has over 2^n interpretations.

Assignment Project Exam Help

“I saw the man with the telescope”: 2 parses

“I saw the man on the hill with the telescope”: 3 parses

“I saw the man on the hill in Texas with the telescope”: 14 parses

“I saw the man on the hill in Texas with the telescope at noon.”: 42 parses

“I saw the man on the hill in Texas with the telescope at noon on Monday” 132 parses

Ambiguity is Ubiquitous

Speech Recognition

- “recognize speech” vs. “wreck a nice beach”
- “youth in Asia” vs. “out of Asia”

Assignment Project Exam Help

Syntactic Analysis

- “I ate spaghetti with chopsticks” vs. “I ate spaghetti with meatballs.”

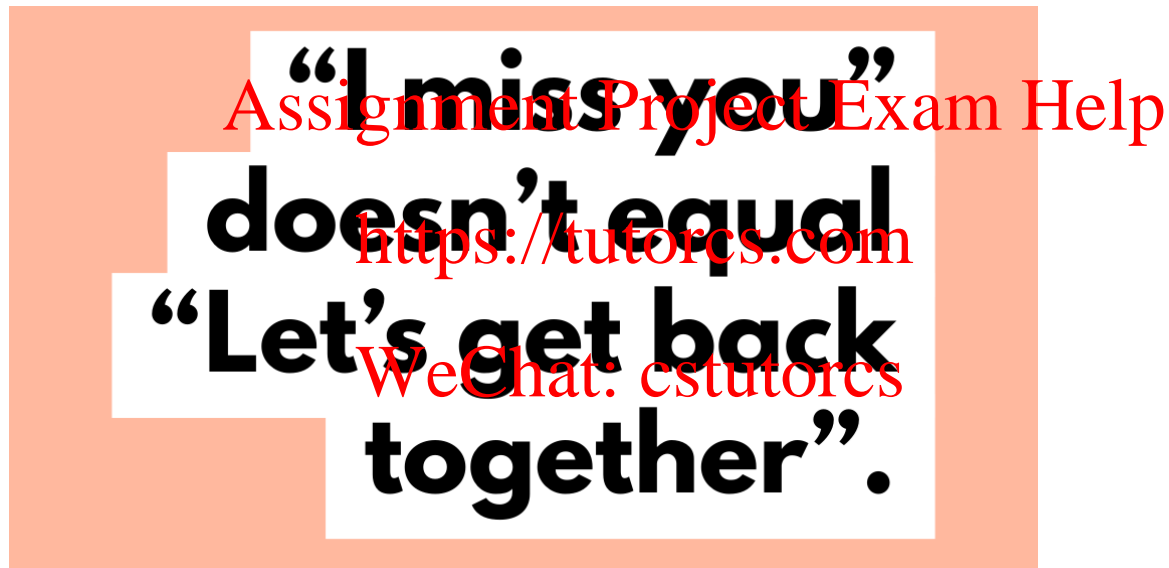
WeChat: cstutorcs

<https://tutorcs.com>

Semantic Analysis

- “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”

Even human struggle with understanding



??????

The difficulty level in various NLP tasks

Easy

- Spell Checking
- Keyword Search
- Finding Synonyms

Assignment Project Exam Help

<https://tutorcs.com>

Medium

- Extracting Information from documents (including websites)

WeChat: cstutorcs

Difficult

- Semantic Analysis (What is the meaning of query statement?)
- Machine Translation
- Coreference Resolution
- Question Answering

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. **Word Meaning and Representation**
4. Count-based Word Representation
 - One-hot Encoding
 - Bag of Words
 - Term Frequency-Inverse Document Frequency
5. Next Week Preview

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

How to represent the meaning of the word?

Definition: meaning (Collins dictionary).

- the idea that it represents, and which can be explained using other words.
- the thoughts or ideas that are intended to be expressed by it.

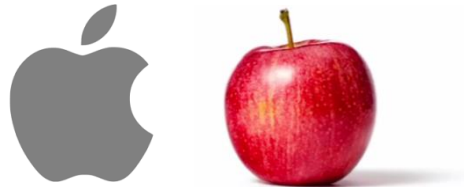
Assignment Project Exam Help

<https://tutorcs.com>

signifier (symbol) \Leftrightarrow signified (idea or thing)

WeChat: cstutores

“Apple”



How do we have usable meaning in a computer?

- Common solution: Use e.g. WordNet, a thesaurus containing lists of synonym sets and hypernyms (“is a” relationships).
- <http://wordnetweb.princeton.edu/perl/webwn>

Assignment Project Exam Help

e.g. synonym sets containing “good”:

e.g. hypernyms of “panda”:

```
from nltk.corpus import wordnet as wn
poses = { 'n': 'noun', 'v': 'verb', 's': 'adj (s)',
          'a': 'adj', 'r': 'adv' }
for synset in wn.synsets("good"):
    print("{}: {}".format(poses[synset.pos()],
                          ", ".join([l.name() for l in synset.lemmas()])))
```

```
noun: good
noun: good, goodness
noun: good, goodness
noun: commodity, trade_good, good
adj: good
adj (sat): full, good
adj: good
adj (sat): estimable, good, honorable, respectable
adj (sat): beneficial, good
adj (sat): good
adj (sat): good, just, upright
...
adverb: well, good
adverb: thoroughly, soundly, good
```

```
from nltk.corpus import wordnet as wn
panda = wn.synset("panda.n.01")
hyper = lambda s: s.hypernyms()
list(panda.closure(hyper))
```

```
[Synset('procyonid.n.01'),
Synset('carnivore.n.01'),
Synset('placental.n.01'),
Synset('mammal.n.01'),
Synset('vertebrate.n.01'),
Synset('chordate.n.01'),
Synset('animal.n.01'),
Synset('organism.n.01'),
Synset('living_thing.n.01'),
Synset('whole.n.02'),
Synset('object.n.01'),
Synset('physical_entity.n.01'),
Synset('entity.n.01')]
```

<https://tutorcs.com>
WeChat: cstutores

Problems with resources like WordNet

- Great as a resource but missing nuance
 - e.g. “proficient” is listed as a synonym for “good”. This is only correct in some contexts.
 - e.g. “glad” can be synonym for “fXXX off”???

Assignment Project Exam Help

- Missing new meanings of words
 - e.g., wicked, badass, nifty, wizard, genius, ninja, bombast
 - Impossible to keep up-to-date!

<https://tutorcs.com>

- Subjective WeChat: cstutorcs
- Requires human labor to create and adapt
- Can’t compute accurate word similarity

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. Word Meaning and Representation
4. **Count-based Word Representation**
 - One-hot Encoding
 - Bag of Words
 - Term Frequency-Inverse Document Frequency
5. Next Week Preview

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

One-hot Encoding

- In traditional NLP, we regard words as discrete symbols.

Hot (True) Cold (False)

Means one 1, the rest 0s

Assignment Project Exam Help

Words can be represented by **one-hot vectors**:

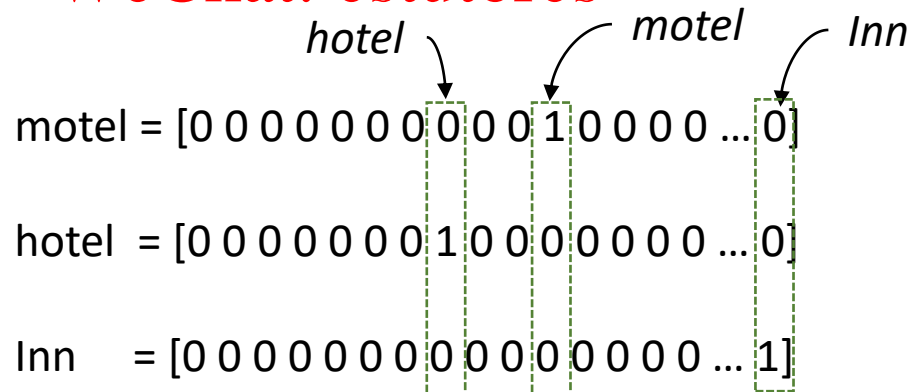
- The categorical values be mapped to integer values (index)
- each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

hotel *motel* *Inn*

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 ... 0]

hotel = [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ... 0]

Inn = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... 1]



Vector dimension = number of words in vocabulary

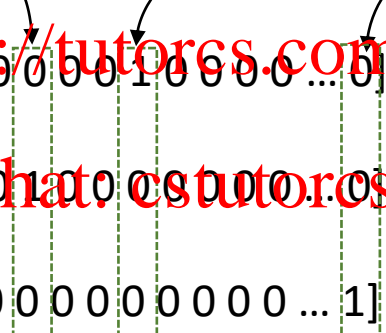
Problem with one-hot vectors

Problem #1. No word similarity representation

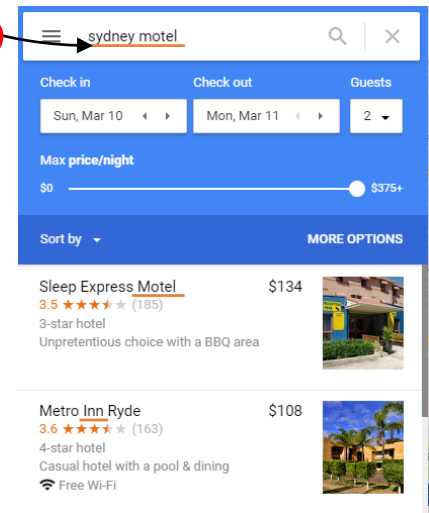
Example: in web search, if user searches for “Sydney motel”, we would like to match documents containing “Sydney Inn”

hotel
motel
Inn

$\text{motel} = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \dots 0]$
 $\text{hotel} = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \dots 0]$
 $\text{Inn} = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \dots 1]$



There is no natural notion of similarity for one-hot vectors!



Problem #2. Inefficiency

Vector dimension = number of words in vocabulary

Each representation has only a single '1' with all remaining 0s.

Bag of Words (BOW)

- A bag-of-words model (BoW) is a representation of text that describes **the occurrence of words** within a document. It involves two things:
 - A vocabulary of known words.
 - A measure of the presence of known words.
- It is called a “**bag**” of words, because any information about the **order or structure of words in the document is discarded**. The model is only concerned with whether known words occur in the document, not where in the document.

How are you
How have you been
Nice to see you
Have a nice day



Bag of Words (BOW)

- A bag-of-words model (BoW) is a representation of text that describes **the occurrence of words** within a document. It involves two things:
 - A vocabulary of known words.
 - A measure of the presence of known words.
- It is called a “bag” of words, because any information about the **order or structure of words in the document is discarded**

Assignment Project Exam Help

<https://tutores.com>

WeChat: cstutorcs

How are you
How have you been
Nice to see you
Have a nice day



Bag of Words (BOW)



A vocabulary of known words

a are been day have how nice see to you

* WO = occurrence of words

[WO_a, WO_{are}, WO_{been}, WO_{day}, WO_{have}, WO_{how}, WO_{nice}, WO_{see}, WO_{to}, WO_{you}]

<https://tutorcs.com>

How are you = [0, 1, 0, 0, 0, 1, 0, 0, 0, 1]

How have you been = [0, 0, 1, 0, 1, 1, 0, 0, 0, 1]

Nice to see you = [0, 0, 0, 0, 0, 0, 1, 1, 1, 1]

Have a nice day = [1, 0, 0, 1, 1, 0, 1, 0, 0, 0]

Total Frequency = [1, 1, 1, 1, 2, 2, 2, 1, 1, 3]

| a | are | been | day | have | how | nice | see | to | you |
|---|-----|------|-----|------|-----|------|-----|----|-----|
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 3 |

Why use BoW?

- The intuition is that documents are similar if they have similar content. Further, that from the content alone we can learn something about the meaning of the document.

Assignment Project Exam Help

Problem with BoW

- Discarding word order ignores the context, and in turn meaning of words in the document (semantics). Context and meaning can offer a lot to the model, that if modeled could tell the difference between the same words differently arranged (“this is interesting” vs “is this interesting”).

S1= I love you but you hate me

S2= I hate you but you love me



Term Frequency-Inverse Document Frequency

- Term Frequency-Inverse Document Frequency (TF-IDF) is a way of representing *how important a word is to a document in a collection or corpus.*

Assignment Project Exam Help

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

<https://tutorcs.com>

$w_{i,j}$ = weight of term i in document j

$tf_{i,j}$ = number of occurrences of term i in document j

WeChat: cstutorcs

N = total number of documents

df_i = number of documents containing term i

- The **Term Frequency** is a count of how many times a word occurs in a given document (synonymous with bag of words)
- The **Document Frequency** is the number of times a word occurs in a corpus of documents

Term Frequency

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

Assignment Project Exam Help

Like BoW

$tf_{i,j}$ = number of occurrences of term i in document j

<https://tutorcs.com>

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

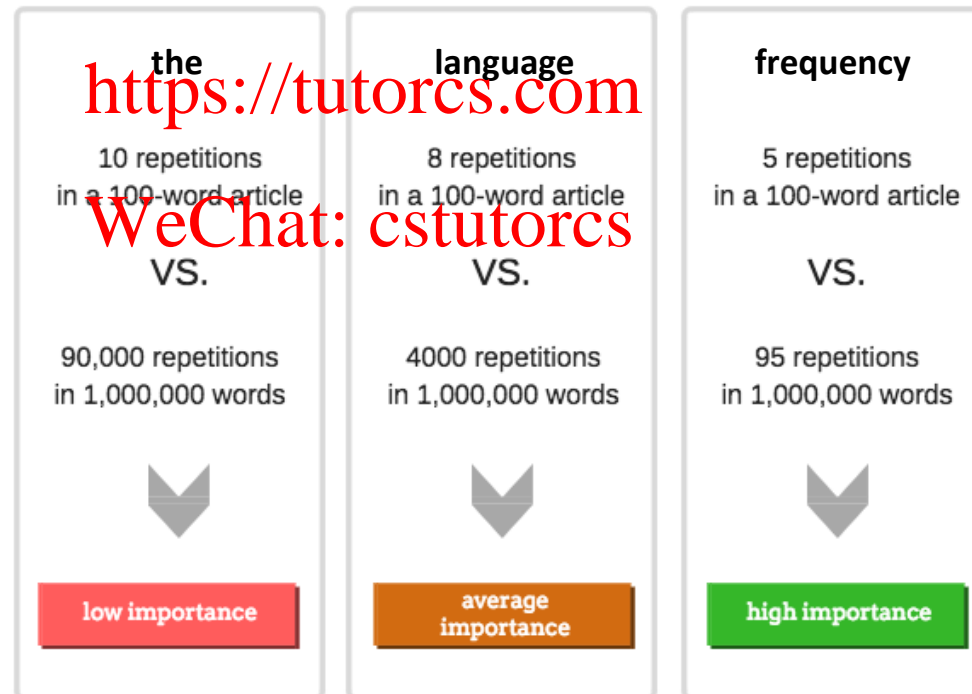
WeChat: cstutorcs

| | and | apple | banana | fruit | I | like | sweet | yellow |
|-----|-----|-------|--------|-------|---|------|-------|--------|
| D#1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| D#2 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| D#3 | 1 | 0 | 2 | 0 | 0 | 0 | 1 | 1 |
| D#4 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

What if we just use Term Frequency Only?

- It is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones.

Assignment Project Exam Help

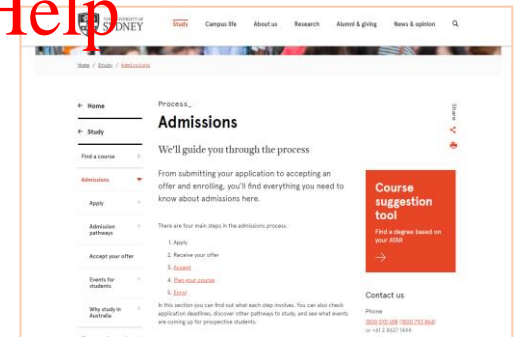




Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutors



Webpage#1

Webpage#2

Webpage#3

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

Why do we need log?

Assignment Project Exam Help

N = total number of documents

df_i = number of documents containing term i

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

WeChat: cstutorcs

$N = 4$

| | and | apple | banana | fruit | I | like | sweet | yellow |
|----|-----|-------|--------|-------|---|------|-------|--------|
| df | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 |

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Why do we need log?

Assignment Project Exam Help

N = total number of documents

df_i = number of documents containing term i

With log $n = 1,000,000$
 $idf(d, t) = \log(n/df(t))$

| | $df(t)$ | $idf(d, t)$ |
|-------|-----------|-------------|
| word1 | 1 | 6 |
| word2 | 100 | 4 |
| word3 | 1,000 | 3 |
| word4 | 10,000 | 2 |
| word5 | 100,000 | 1 |
| word6 | 1,000,000 | 0 |

Without log
 $idf(d, t) = n/df(t)$

| | $df(t)$ | $idf(d, t)$ |
|-------|-----------|-------------|
| word1 | 1 | 1,000,000 |
| word2 | 100 | 10,000 |
| word3 | 1,000 | 1,000 |
| word4 | 10,000 | 100 |
| word5 | 100,000 | 10 |
| word6 | 1,000,000 | 1 |

WeChat: cstutorcs

<https://tutors.com>

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

Assignment Project Exam Help

$1+df_i$ sometimes, why?

N = total number of documents

df_i = number of documents containing term i

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

WeChat: cstutorcs

$N = 4$

| | and | apple | banana | fruit | I | like | sweet | yellow |
|----|-----|-------|--------|-------|---|------|-------|--------|
| df | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 |

Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)^{1+df_i}$$

Assignment Project Exam Help

N = total number of documents

df_i = number of documents containing term i

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

WeChat: cstutorcs

$N = 4$

| | and | apple | banana | fruit | I | like | sweet | yellow |
|-------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| df | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 |
| idf (with $1+df_i$) | $\text{Inv}(4/(1+1))$ =0.693147 | $\text{Inv}(4/(1+1))$ =0.693147 | $\text{Inv}(4/(2+1))$ =0.287682 | $\text{Inv}(4/(1+1))$ =0.693147 | $\text{Inv}(4/(2+1))$ =0.287682 | $\text{Inv}(4/(2+1))$ =0.287682 | $\text{Inv}(4/(2+1))$ =0.287682 | $\text{Inv}(4/(1+1))$ =0.693147 |

We use a natural logarithm to the base of the mathematical constant e .

where e is an irrational and transcendental number approximately equal to 2.718281828459

Term Frequency Inverse Document Frequency

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)^{1+df_i}$$

Assignment Project Exam Help

$w_{i,j}$ = weight of term i in document j

<https://tutorcs.com>

Document #1: I like apple

Document #2: I like banana

Document #3: Sweet and yellow banana banana

Document #4: Sweet fruit

WeChat: cstutorcs

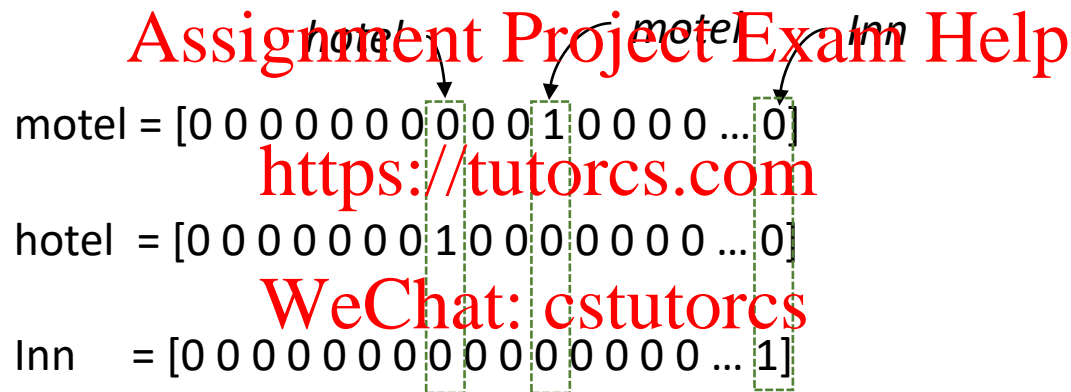
| | and | apple | banana | fruit | I | like | sweet | yellow |
|-----|----------|----------|----------|----------|----------|----------|----------|----------|
| D#1 | 0 | 0.693147 | 0 | 0 | 0.287682 | 0.287682 | 0 | 0 |
| D#2 | 0 | 0 | 0.287682 | 0 | 0.287682 | 0.287682 | 0 | 0 |
| D#3 | 0.693147 | 0 | 0.575364 | 0 | 0 | 0 | 0.287682 | 0.693147 |
| D#4 | 0 | 0 | 0 | 0.693147 | 0 | 0 | 0.287682 | 0 |

Sparse Representation

With COUNT based word representation (especially, one-hot vector), linguistic information was represented with sparse representations (high-dimensional features)

Assignment Project Exam Help
<https://tutorcs.com>
WeChat: cstutorcs

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 ... 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ... 0]
Inn = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... 1]



However, with the recent popularity and success of **word embeddings (low dimensional, distributed representations)**, **neural-based models have achieved superior results on various language-related tasks** as compared to traditional machine learning models with high-dimensional features.

Lecture 1: Introduction to Natural Language Processing

1. Course Introduction
2. Overview of Natural Language Processing (NLP)
3. Word Meaning and Representation
4. **Count-based Word Representation**
 - One-hot Encoding
 - Bag of Words
 - Term Frequency-Inverse Document Frequency
5. Next Week Preview

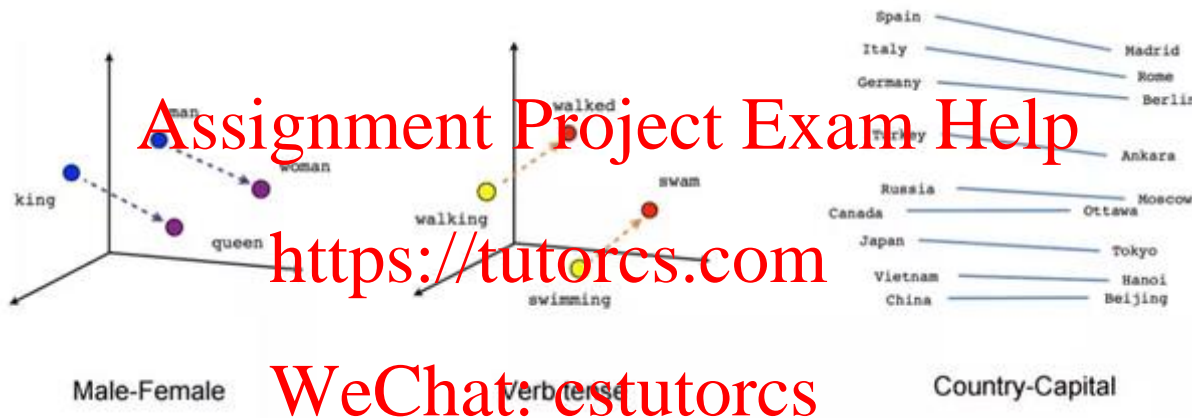
Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

How to Represent the Word Similarity!

- How to represent the word similarity with dense vector



- Try this with word2vec

Word Algebra

Enter all three words, the first two, or the last two and see the words that result.

shanghai + (australia - sydney) = 0.7477672216910414

Reference for this lecture

- Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.
- Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. "O'Reilly Media, Inc".
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Manning, C 2017, Introduction and Word Vectors, Natural Language Processing with Deep Learning, lecture notes, Stanford University
- Mooney, RJ 2000, Natural Language Processing Introduction, CS388, lecture notes, University of Texas at Austin

Assignment Project Exam Help

<https://tutorcs.com>

WeChat: cstutorcs

Welcome to Hell! ♥

Assignment Project Exam Help

<https://tutores.com>

WeChat: cstutores