DIGITAL TALENT SCHOLARSHIP

KOMINFO

DIDUKUNG OLEH :
Microsoft

Penambangan Teks dan NLP

Supeno Mardi

Daftar isi

- Definisi Text Mining
- Text Analytic
- Text Mining dengan NLP
- Proses Text Mining
- NLP

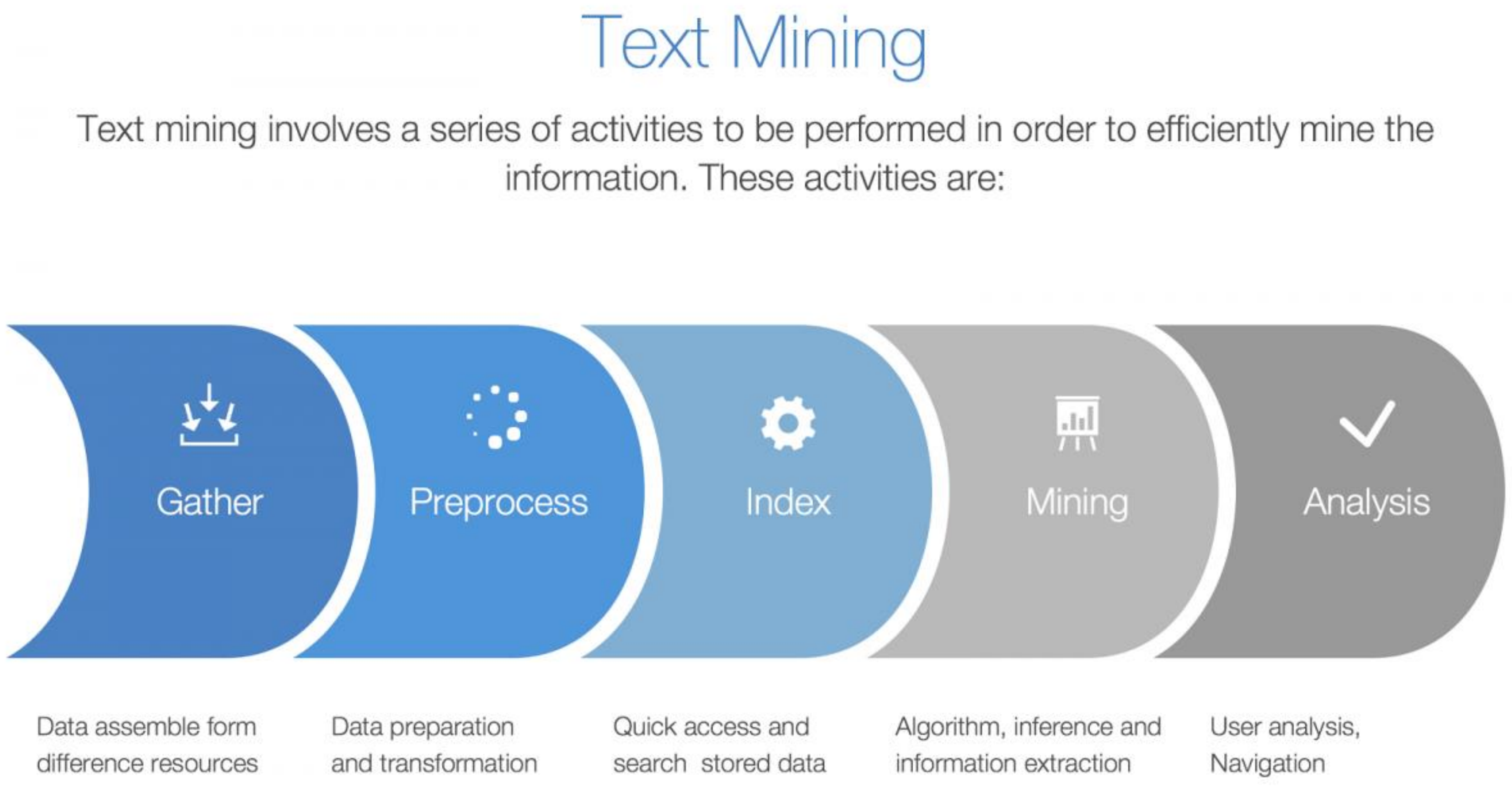
TEXT MINING

- Penambangan teks berusaha untuk mengekstrak informasi yang berguna dan penting dari dokumen berformat heterogen, seperti halaman web, email, posting media sosial, artikel jurnal, dll.

Kegunaan Text Mining

- Research: *e.g. knowledge discovery, medical/healthcare* - di masa lalu butuh banyak waktu bagi peneliti manusia untuk menganalisis dan memperoleh informasi yang relevan. Dalam beberapa kasus, informasi ini bahkan tidak dapat diambil. Penambangan teks memungkinkan peneliti menemukan informasi lebih banyak dan dengan cara yang lebih cepat dan lebih efisien.
- Business: e.g. *risk management*, *resume filtering* - Perusahaan besar menggunakan penambangan teks untuk membantu pengambilan keputusan dan untuk menjawab pertanyaan pelanggan dengan cepat.
- Security: *e.g. anti-terrorism* - analisis blog dan sumber teks online lainnya digunakan untuk mencegah kejahatan internet dan melawan penipuan.
- Daily, *e.g. spam filtering*, *social media data analysis* - penambangan teks digunakan oleh situs web email untuk menciptakan metode penyaringan yang lebih andal dan efektif. Ini juga digunakan untuk tujuan media sosial dengan mengidentifikasi hubungan antara pengguna dan produk tertentu atau untuk menentukan pendapat pengguna tentang topik tertentu

Langkah-langkah Text Mining



Text Mining

Text mining involves a series of activities to be performed in order to efficiently mine the information. These activities are:

- **Gather**: Data assemble form difference resources
- **Preprocess**: Data preparation and transformation
- **Index**: Quick access and search stored data
- **Mining**: Algorithm, inference and information extraction
- **Analysis**: User analysis, Navigation

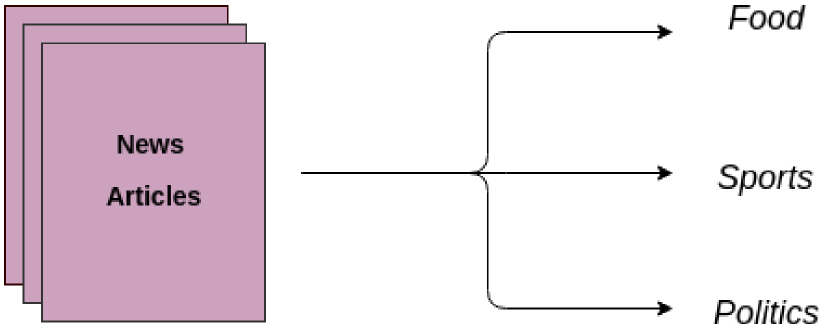
ANALISA TEKS/TEXT **ANALYTICS**

- Analisis teks membantu analis mengekstraksi makna, pola, dan struktur yang tersembunyi dalam data tekstual yang tidak terstruktur.
- Analisis teks menggabungkan alat dan teknik yang digunakan untuk memperoleh wawasan dari data yang tidak terstruktur

TEKNIK-TEKNIK TEXT ANALYTICS

- information retrieval
  - ilmu mencari informasi dalam dokumen, mencari dokumen sendiri, dan juga mencari metadata yang menggambarkan data, dan untuk database teks, gambar atau suara.
- exploratory analysis
  - adalah pendekatan untuk menganalisis set data untuk meringkas karakteristik utama mereka, seringkali dengan metode visual
- concept extraction
  - Ekstraksi konsep adalah teknik menambang topik yang paling penting dari sebuah dokumen. Dalam konteks e-commerce, ekstraksi konsep dapat digunakan untuk mengidentifikasi apa yang terkait dengan halaman belanja Web
- Summarization
  - proses pemendekan dokumen teks dengan perangkat lunak, untuk membuat ringkasan dengan poin-poin utama dari dokumen asli

- Categorization
- Sentiment analysis
- Content management
- Ontology management

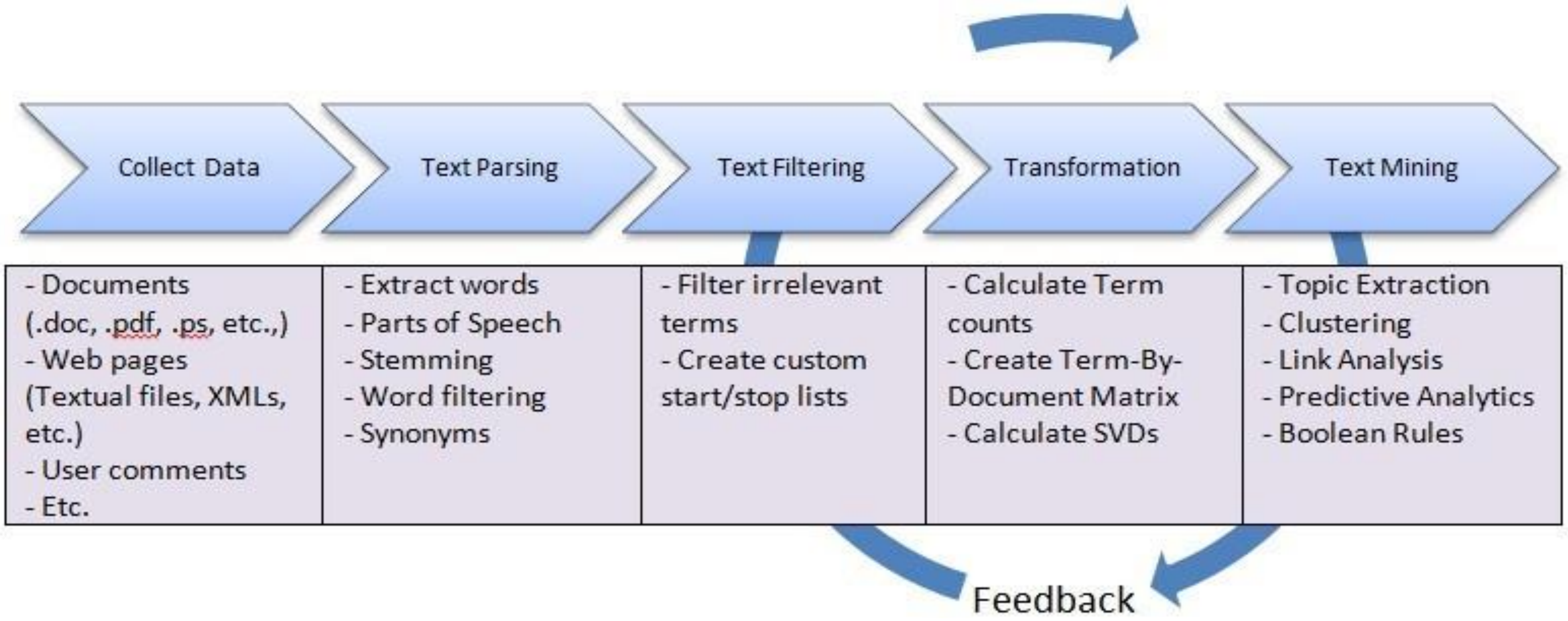


<table>
<tr><th colspan="5">Text Analytics</th></tr>
<tr><td colspan="3">Search
(Information Organization and Access)</td><td colspan="2">Descriptive and Predictive Analysis (Discovering Trends, Patterns, and Modeling)</td></tr>
<tr><td>Information Retrieval</td><td>Content Categorization</td><td>Ontology Management</td><td>Text Mining</td><td>Sentiment Analysis</td></tr>
</table>

Penambangan Teks dengan NLP

- Data-mining: Ekstraksi informasi (atau pola) yang menarik dari data terstruktur.
- Untuk Penambangan Teks, data-mining berbasis teknik Natural Language Processing (NLP)
- Text Mining = *Statistical NLP* (structured data) + *Data mining* (pattern discovery)

Proses Penambangan Teks



Collect Data	Text Parsing	Text Filtering	Transformation	Text Mining
- Documents (.doc, .pdf, .ps, etc.,) - Web pages (Textual files, XMLs, etc.) - User comments - Etc.	- Extract words - Parts of Speech - Stemming - Word filtering - Synonyms	- Filter irrelevant terms - Create custom start/stop lists	- Calculate Term counts - Create Term-By-Document Matrix - Calculate SVDs	- Topic Extraction - Clustering - Link Analysis - Predictive Analytics - Boolean Rules

Proses Penambangan Teks

- Text Preprocessing
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Data Mining
  - Classification (Supervised) / Clustering (Unsupervised)
- Analyzing results

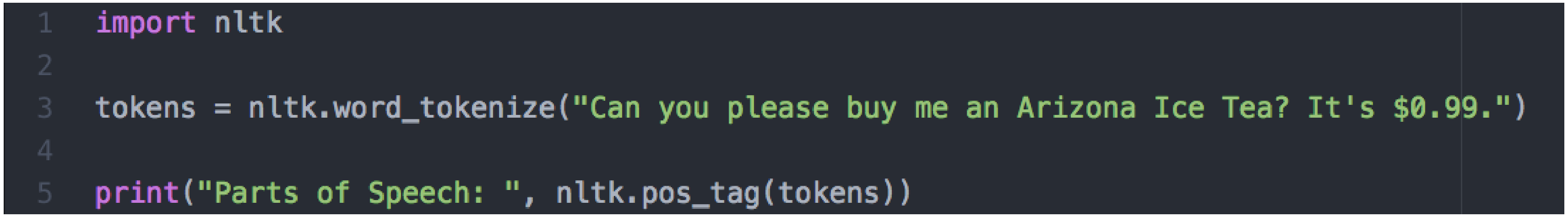
Proses Penambangan Teks

- Text Preprocessing
  - *Part Of Speech (POS)* tagging
    - Find the corresponding POS for each word.
  - Word sense *disambiguation*
    - Context based or proximity based
  - *Parsing*
    - Generates a parse tree for each sentence

POS Tagging

- Can you please buy me an Arizona Ice Tea? It's $0.99.

```
[‘Can’, ‘you’, ‘please’, ‘buy’, ‘me’, ‘an’,‘Arizona’,
‘Ice’, ‘Tea’, ‘?’, ‘It’, “‘s”, ‘$’, ‘0.99’, ‘.’]
```



```
import nltk

tokens = nltk.word_tokenize("Can you please buy me an Arizona Ice Tea? It's $0.99.")

print("Parts of Speech: ", nltk.pos_tag(tokens))
```

```
Part of speech  [('Can', 'MD'), ('you', 'PRP’),
('please', 'VB'), ('buy', 'VB'), ('me', 'PRP'),
('an', 'DT'), ('Arizona', 'NNP'), ('Ice', 'NNP'),
('Tea', 'NNP'), ('?', '.'), ('It', 'PRP'), ("'s",
'VBZ'), ('$', '$'), ('0.99', 'CD'), ('.', '.')]
```

Daftar dari POS tagging

- CC coordinating conjunction
- CD cardinal digit
- DT determiner
- EX existential there (like: “there is” … think of it like “there exists”)
- FW foreign word
- IN preposition/subordinating conjunction
- JJ adjective ‘big’
- JJR adjective, comparative ‘bigger’
- JJS adjective, superlative ‘biggest’
- LS list marker 1)
- MD modal could, will
- NN noun, singular ‘desk’
- NNS noun plural ‘desks’
- NNP proper noun, singular ‘Harrison’
- NNPS proper noun, plural ‘Americans’
- PDT predeterminer ‘all the kids’
- POS possessive ending parent’s
- PRP personal pronoun I, he, she
- PRP$ possessive pronoun my, his, hers
- RB adverb very, silently,
- RBR adverb, comparative better
- RBS adverb, superlative best
- RP particle give up
- TO, to go ‘to’ the store.
- UH interjection, errrrrrrrm
- VB verb, base form take
- VBD verb, past tense took
- VBG verb, gerund/present participle taking
- VBN verb, past participle taken
- VBP verb, sing. present, non-3d take
- VBZ verb, 3rd person sing. present takes
- WDT wh-determiner which
- WP wh-pronoun who, what
- WP$ possessive wh-pronoun whose
- WRB wh-abverb where, when

LATIHAN

- Install NLTK

```
pip install nltk

>>> import nltk
>>> nltk.download()
```

Select All Package!

- Lalu coba bermacam kalimat sbb:
  - Hello, my name is…
  - Would you like a tea?
  - Where is your address?

Proses Penambangan Teks

- Feature Generation
  - Text document is represented by the words it contains (and their occurrences)
    - Order of words is not that important for certain applications (Bag of words)
  - *Stemming*: identifies a word by its root
    - Reduce dimensionality
  - Stop words: The common words unlikely to help text mining

Bag of Words

- ‘All my cats in a row’,‘When my cat sits down, she looks like a Furby toy!’,
- {‘all’: 0, ‘cat’: 1, ‘cats’: 2, ‘down’: 3, ‘furby’: 4, ‘in’: 5, ‘like’: 6, ‘looks’: 7, ‘my’: 8, ‘row’: 9, ‘she’: 10, ‘sits’: 11, ‘toy’: 12, ‘when’: 13 }
- ‘All my cats in a row’ = [1 0 1 0 0 1 0 0 1 1 0 0 0 0]

LATIHAN

```
from sklearn.feature_extraction.text import
CountVectorizer

corpus = [
'All my cats in a row',
'When my cat sits down, she looks like a Furby toy!’,
]

vectorizer = CountVectorizer()
print(vectorizer.fit_transform(corpus).todense())
print(vectorizer.vocabulary_)
```

Proses Penambangan Teks

- Feature Selection
  - *Reduce dimensionality*
    - Learners have difficulty addressing tasks with high dimensionality
    - Only interested in the information relevant to what is being analyzed
  - Irrelevant features
    - Not all features help

Proses Penambangan Teks

- Supervised learning (classification)
  - The training data is labeled indicating the class

    New data is classified based on the training set
  - Correct classification: The known label of test sample is identical with the class result from the classification model

- Unsupervised learning (clustering)
  - The class labels of training data are unknown
  - Establish the existence of classes or clusters in the data
  - Good clustering method: high intra-cluster similarity and low inter-cluster similarity

Natural Language Processing

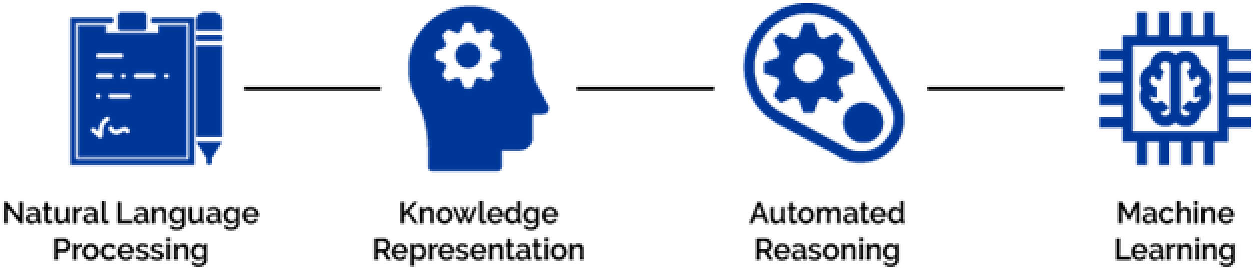
Natural Language Processing (NLP)

Natural Language Understanding (NLU)

Mengerti Bahasa yang ditulis/dikatakan

Natural Language Generation

Mengekspresikan Bahasa secara alami (oleh mesin)



Communication Typical communication episode

S (speaker) ingin menyampaikan P (proposition) kepada H (hearer) menggunakan W (words in a formal or natural language)

1. Speaker

- **Intention:** S ingin H mengetahui P
- **Generation:** S memilih dan merangkai kata-kata W
- **Synthesis:** S mengucapkan/menyampaikan kata-kata W

2. Hearer

- **Perception:** H mempersepsi kata-kata W" (ideally W" = W)
- **Analysis:** H menyimpulkan makna yang mungkin P1,P2,…,Pn untuk W"
- **Disambiguation:** H menyimpulkan bahwa S bermaksud menyampaikan Pi (ideally Pi=P)
- **Incorporation:** H memutuskan tahu atau tidak tahu Pi

Aplikasi-aplikasi NLP

- Machine Translation
- Extracting data from text
    - Konversi unstructured text ke structure data
- Spoken language control systems
- Spelling and grammar checkers

Natural language understanding

Raw speech signal

↓

- **Speech recognition**

Sequence of words spoken

↓

- **Syntactic analysis** using knowledge of the grammar

Structure of the sentence

↓

- **Semantic analysis** using info. about meaning of words

Partial representation of meaning of sentence

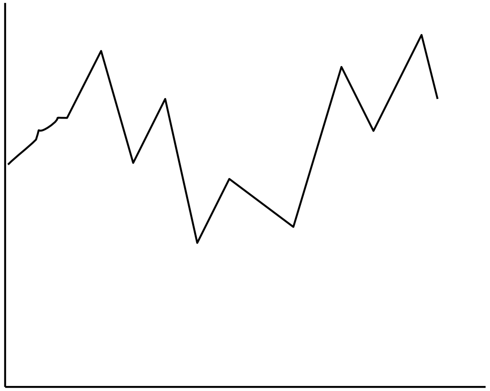
↓

- **Pragmatic analysis** using info. about context

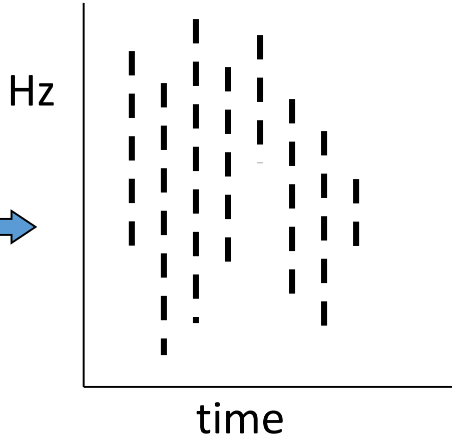
Final representation of meaning of sentence

Speech Recognition

Input (microphone records voice)



Analog Signal



Freq. spectrogram (e.g. Fourier transform)

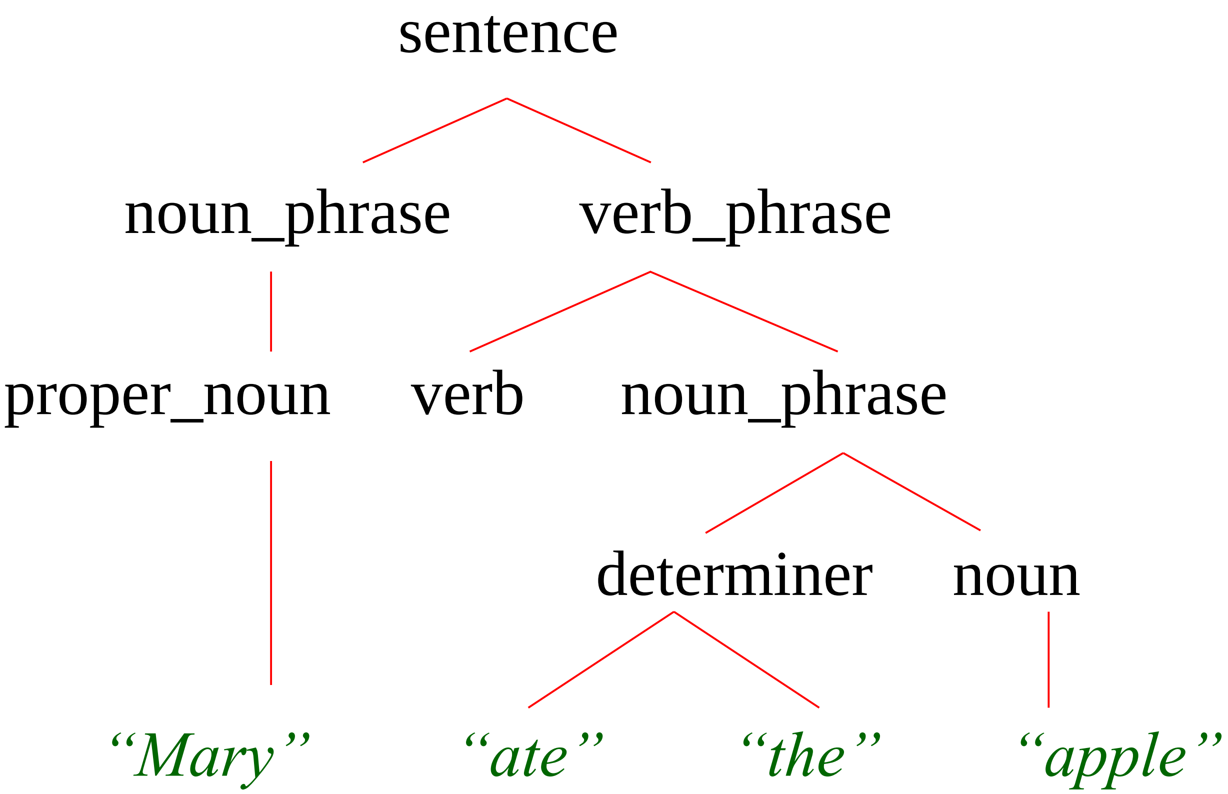
Syntactic Analysis

- Rules of syntax (grammar) specify the possible organization of words in sentences and allows us to determine sentence’s structure(s)
  - “John saw Mary with a telescope”
    - John saw (Mary with a telescope)
    - John (saw Mary with a telescope)
- Parsing: given a sentence and a grammar
  - Checks that the sentence is correct according with the grammar and if so returns a **parse tree** representing the structure of the sentence

Syntactic Analysis - Grammar

- sentence -> noun_phrase, verb_phrase
- noun_phrase -> proper_noun
- noun_phrase -> determiner, noun
- verb_phrase -> verb, noun_phrase
- proper_noun -> [mary]
- noun -> [apple]
- verb -> [ate]
- determiner -> [the]

Syntactic Analysis - Parsing



sentence

noun_phrase verb_phrase

proper_noun verb noun_phrase

determiner noun

"Mary" *"ate"* *"the"* *"apple"*

Semantic Analysis

- Generates (partial) meaning/representation of the sentence from its syntactic structure(s)
- Compositional semantics: meaning of the sentence from the meaning of its parts:
  - Sentence: A tall man likes Mary
  - Representation: man(x) & tall(x) & likes(x, mary)
- Grammar + Semantics
  - `Sentence (Smeaning)-> noun_phrase(NPmeaning),verb_phrase(VPmeaning), combine(NPmeaning,VPmeaning,Smeaning)`