# Natural Language Processing

Abdelhak Mahmoudi
abdelhak.mahmoudi@um5.ac.ma

Direction Générale des Impôts

February 5st, 2020

# Content

1. Regular Expressions
2. Tokenization
3. Character Encoding
4. Part-of-Speech Tagging
5. Chunking
6. Stemming and Lemmatization
7. Parsing
8. Named Entity Recognition
9. Topic Segmentation

# Regular Expressions

- Meta Characters: Character matches
  - . : wildcard, match a single character
  - ^: start of a string
  - $: end of a string
  - []: matches one of the set of characters within []
  - []a-z: matches one of the range of characters a, b, …z
  - [^abc]: matches a character that is not a, b or c.
  - aIb: matches either a or b, where a and b are strings
  - (): scopeing of operators
  - \: escape character for special characters (\t, \n, \b)

# Regular Expressions

- Meta Characters: Character Symbols
  - \b: matches word boundary
  - \d: any digit, equivalent to [0-9]
  - \D: any non-digit, equivalent to [^0-9]
  - \s: any whitespace, equivalent to [\t\n\r\f\v]
  - \S: any non whitespace, equivalent to [^\t\n\r\f\v]
  - \w: alpha-numeric character, equivalent to [a-zA-Z0-9_]
  - \W: non alpha-numeric character , equivalent to [^a-zA-Z0-9_]

# Regular Expressions

- Meta Characters: Repetitions
  - *: matches zero or more occurrences
  - +: matches one or more occurrences
  - ?: matches zero or one occurrences
  - {n}: exactly n repetitions, n>=0
  - {n,}: at least n repetitions
  - {,n}: at most n repetitions
  - {m,n}: at least m and at most m repetitions

# Regular Expressions

- Examples
  - *Dates*

"5-2-2020, 15/2/2020, 2020/2/4 autre autre"

r'(\d{1,4}[.\-/]\d{1,2}[.\-/]\d{1,4})'

  - *Emails*

"ahmed@dgi.gov.ma, maryam@dgi.ma ahmadi3maryam@gmail.com other text here"

r'[\w.-]+@[\w.-]+'

# Character Encoding

Monsieur le directeur, j'ai le plaisir de vous informer d'un cas d'evasion fiscale concernant la société SOKA dirigé par Monsieur Ahmadi Ahmed. La société est domiciliée à l'adresse 31 Boulevard ANNASR, Rabat, Maroc. Son registre de commerce RC: 112233 et ICE: 445566778899. Veuillez agréer Monsieur, mes salutations distinguées.

السيد المدير ، يسرني أن أبلغكم بحالة التهرب الضريبي فيما يتعلق بشركة SOKA برئاسة السيد أحمدي أحمد. يقع مقر الشركة في 31 شارع النصر، الرباط ، المغرب. السجل التجاري RC: 112233 و ICE: 445566778899. تفضلوا سيدي بقبول أطيب تحياتي.

Arabic text encoding

- Windows-1256,
- UTF-8,
- CP720,
- ISO 8859-6.

Arabic text display Reshaper

- Unicode bidirectional algorithm, implemented with python-bidi.
- http://pydj.mpcabd.xyz/arabic-reshaper/
- https://camel.abudhabi.nyu.edu/madamira/

# Tokenization

- Detect patterns in text
- Word Tokenization

['Monsieur', 'le', 'directeur', ',', "j'ai", 'le', 'plaisir', 'de', 'vous', 'informer', "d'un", 'cas', "d'evasion", 'fiscale', 'concernant', 'la', 'société', 'SOKA', 'dirigé', 'par', 'Monsieur', 'Ahmadi', 'Ahmed', '.', 'La', 'société', 'est', 'domiciliée', 'à', "l'adresse", '31', 'Boulevard', 'ANNASR', ',', 'Rabat', ',', 'Maroc', '.', 'Son', 'registre', 'de', 'commerce', 'RC', ':', '112233', 'et', 'ICE', ':', '445566778899', '.', 'Veuillez', 'agréer', 'Monsieur', ',', 'mes', 'salutations', 'distinguées', '.']

# Tokenization

- Detect patterns in text
- Word Tokenization
- Sentence Tokenization

["Monsieur le directeur, j'ai le plaisir de vous informer d'un cas d'evasion fiscale concernant la société SOKA dirigé par Monsieur Ahmadi Ahmed.", "La société est domiciliée à l'adresse 31 Boulevard ANNASR, Rabat, Maroc.",

'Son registre de commerce RC: 112233 et ICE: 445566778899.',
'Veuillez agréer Monsieur, mes salutations distinguées.']

# POS tagging

| Abbreviation | Meaning |
|---|---|
| CC | coordinating conjunction |
| CD | cardinal digit |
| DT | determiner |
| EX | existential there |
| FW | foreign word |
| IN | preposition/subordinating conjunction |
| JJ | adjective (large) |
| JJR | adjective, comparative (larger) |
| JJS | adjective, superlative (largest) |
| LS | list market |
| MD | modal (could, will) |
| NN | noun, singular (cat, tree) |
| NNS | noun plural (desks) |
| NNP | proper noun, singular (sarah) |
| NNPS | proper noun, plural (indians or americans) |
| PDT | predeterminer (all, both, half) |
| POS | possessive ending (parent\'s) |
| PRP | personal pronoun (hers, herself, him,himself) |
| PRP$ | possessive pronoun (her, his, mine, my, our ) |

| Abbreviation | Meaning |
|---|---|
| RB | adverb (occasionally, swiftly) |
| RBR | adverb, comparative (greater) |
| RBS | adverb, superlative (biggest) |
| RP | particle (about) |
| TO | infinite marker (to) |
| UH | interjection (goodbye) |
| VB | verb (ask) |
| VBG | verb gerund (judging) |
| VBD | verb past tense (pleaded) |
| VBN | verb past participle (reunified) |
| VBP | verb, present tense not 3rd person singular(wrap) |
| VBZ | verb, present tense with 3rd person singular (bases) |
| WDT | wh-determiner (that, what) |
| WP | wh- pronoun (who) |
| WRB | wh- adverb (how) |

[('Monsieur', 'NNP'), ('le', 'CC'), ('directeur,', 'JJ'), ("j'ai", 'NN'), ('le', 'NN'), ('plaisir', 'NN'), ('de', 'IN'), ('vous', 'JJ'), ('informer', 'NN'), ("d'un", 'NN'), ('cas', 'NN'), ("d'evasion", 'NN'), ('fiscale', 'NN'), ('concernant', 'NN'), ('la', 'NN'), ('société', 'FW'), ('SOKA', 'NNP'), ('dirigé', 'NN'), ('par', 'NN'), ('Monsieur', 'NNP'), ('Ahmadi', 'NNP'), ('Ahmed.', 'NNP'), ('La', 'NNP'), ('société', 'NN'), ('est', 'JJS'), ('domiciliée', 'NN'), ('à', 'NNP'), ("l'adresse", 'VBZ'), ('31', 'CD'), ('Boulevard', 'NNP'), ('ANNASR,', 'NNP'), ('Rabat,', 'NNP'), ('Maroc.', 'NNP'), ('Son', 'NNP'), ('registre', 'FW'), ('de', 'FW'), ('commerce', 'NN'), ('RC:', 'NNP'), ('112233', 'CD'), ('et', 'NN'), ('ICE:', 'NNP'), ('445566778899.', 'CD'), ('Veuillez', 'NNP'), ('agréer', 'NN'), ('Monsieur,', 'NNP'), ('mes', 'VBZ'), ('salutations', 'NNS'), ('distinguées.', 'NN')]

# Chunking

Regular Expressions in POS taggers

Example: Search for
<NNP.?>*<CD.?>

RC:/NNP 112233/CD
ICE:/NNP 445566778899./CD

| Name of symbol | Description |
|---|---|
| . | Any character except new line |
| * | Match 0 or more repetitions |
| ? | Match 0 or 1 repetitions |

[('Monsieur', 'NNP'), ('le', 'CC'), ('directeur,', 'JJ'), ("j'ai", 'NN'), ('le', 'NN'), ('plaisir', 'NN'), ('de', 'IN'), ('vous', 'JJ'), ('informer', 'NN'), ("d'un", 'NN'), ('cas', 'NN'), ("d'evasion", 'NN'), ('fiscale', 'NN'), ('concernant', 'NN'), ('la', 'NN'), ('société', 'FW'), ('SOKA', 'NNP'), ('dirigé', 'NN'), ('par', 'NN'), ('Monsieur', 'NNP'), ('Ahmadi', 'NNP'), ('Ahmed.', 'NNP'), ('La', 'NNP'), ('société', 'NN'), ('est', 'JJS'), ('domiciliée', 'NN'), ('à', 'NNP'), ("l'adresse", 'VBZ'), ('31', 'CD'), ('Boulevard', 'NNP'), ('ANNASR,', 'NNP'), ('Rabat,', 'NNP'), ('Maroc.', 'NNP'), ('Son', 'NNP'), ('registre', 'FW'), ('de', 'FW'), ('commerce', 'NN'), ('RC:', 'NNP'), ('112233', 'CD'), ('et', 'NN'), ('ICE:', 'NNP'), ('445566778899.', 'CD'), ('Veuillez', 'NNP'), ('agréer', 'NN'), ('Monsieur,', 'NNP'), ('mes', 'VBZ'), ('salutations', 'NNS'), ('distinguées.', 'NN')]

# Stemming and Lemmatization

- Used for text cleaning: map a group of words to the same root form
- Removing the suffixes or prefixes
- Stemming
  - Apply a set of rules to extract the stem (is fast)
  - The stem might not be an actual language word
- Lemmatization
  - The lemma is an actual language word
  - Based on WordNet corpus (is slow)

plaisir ===> plais
de ===> de
vous ===> vous
informer ===> inform

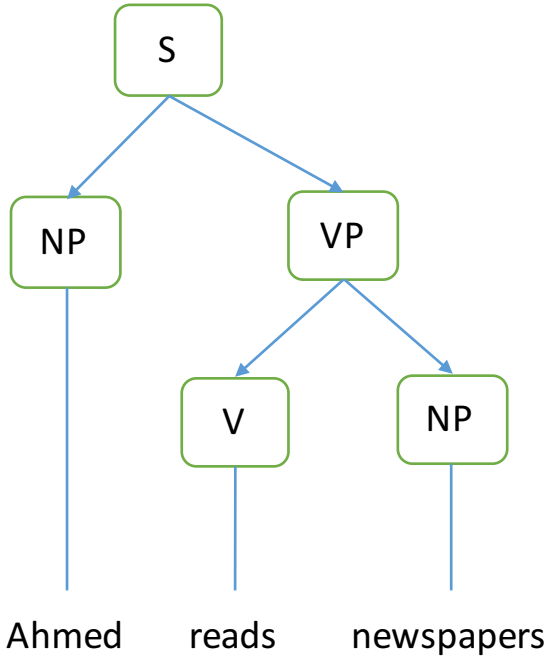
# Parsing

- Parsing: Extract meaning from a sequence of words
- Lexicon : vocabulary of all possible words
- Grammar: how the words have to be linked together

Goal: NP VP
VP: Verb
VP: Verb NP
VP: Verb NP PP
PP: Preposition NP
NP: Noun
NP: Article NP
NP: Adjective Noun
NP: NP PP

"Ahmed reads newspapers"

Noun Verb Noun



# Parsing

- **English**
  - **Treebank:** Large grammar tree learned form wall street Journal database
- **Arabic**
  - Penn Arabic Treebank (ATB): from newswires
- **Complexity**
  - "Une idée verte dors furieusement »

# Named Entity Recognition

- Locating and classifying named entities in texts
- Recognize places, people, dates, values, organizations, etc.

| |
|---|
| 3 class: Location, Person, Organization |
| 4 class: Location, Person, Organization, Misc |
| 7 class: Location, Person, Organization, Money, Percent, Date, Time |

- Language dependent
  - English -> use NLTK
  - French -> train your model
  - Arabic -> train your model
  - https://nlp.stanford.edu/software/CRF-NER.shtml

# Topic Segmentation

- TF-IDF
  - Measures how relevant a term is in a document
  - Map words to vectors
  - Map documents to vectors
  - Ignore the order of words

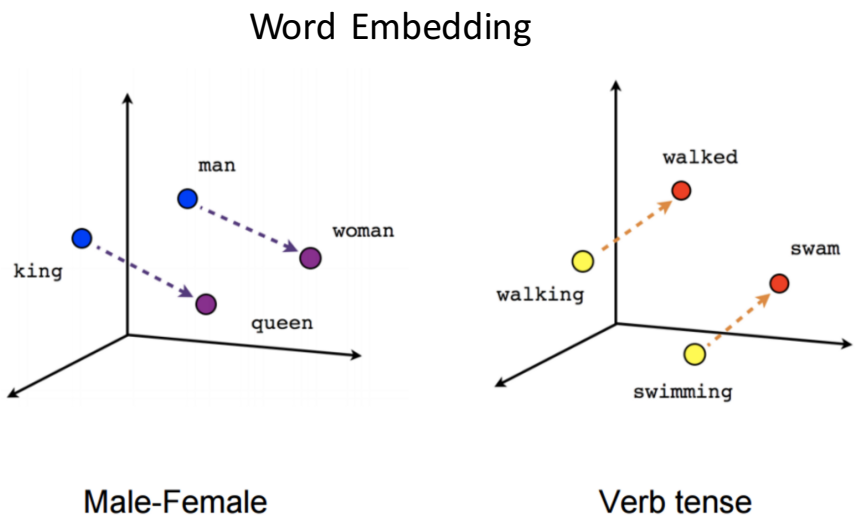
- A : “a new car, used car, car review”
- B : “a friend in need is a friend indeed”

| word | TF | | IDF | TF * IDF | |
|---|---|---|---|---|---|
| | A | B | | A | B |
| a | 1 / 7 | 2 / 8 | Log ( 2 / 2 ) = 0 | 0 | 0 |
| new | 1 / 7 | 0 | Log ( 2 / 1 ) = 0.3 | 0.04 | 0 |
| car | 3 / 7 | 0 | Log ( 2 / 1 ) = 0.3 | 0.13 | 0 |
| used | 1 / 7 | 0 | Log ( 2 / 1 ) = 0.3 | 0.04 | 0 |
| review | 1 / 7 | 0 | Log ( 2 / 1 ) = 0.3 | 0.04 | 0 |
| friend | 0 | 2 / 8 | Log ( 2 / 1 ) = 0.3 | 0 | 0.08 |
| in | 0 | 1 / 8 | Log ( 2 / 1 ) = 0.3 | 0 | 0.04 |
| need | 0 | 1 / 8 | Log ( 2 / 1 ) = 0.3 | 0 | 0.04 |
| is | 0 | 1 / 8 | Log ( 2 / 1 ) = 0.3 | 0 | 0.04 |
| indeed | 0 | 1 / 8 | Log ( 2 / 1 ) = 0.3 | 0 | 0.04 |

# Topic Segmentation

- Word Embedding (Word2Vec)
  - Map words to vectors
  - Take into account word's context and position
  - Cosine Similarity

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$



# Edit Distance

- Levenshtein Distance
  - Measure of similarity between two strings
  - minimum number of edit operations
    - deletions,
    - insertions,
    - substitutions

| | | m | o | n | k | e | y |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| m | 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| o | 2 | 1 | 0 | 1 | 2 | 3 | 4 |
| n | 3 | 2 | 1 | 0 | 1 | 2 | 3 |
| e | 4 | 3 | 2 | 1 | 1 | 1 | 2 |
| y | 5 | 4 | 3 | 2 | 2 | 2 | 1 |

# Tools

- NLTK
  - Documentation https://www.nltk.org/
  - import nltk
  - nltk.download()
- Google API
  - https://cloud.google.com/natural-language/