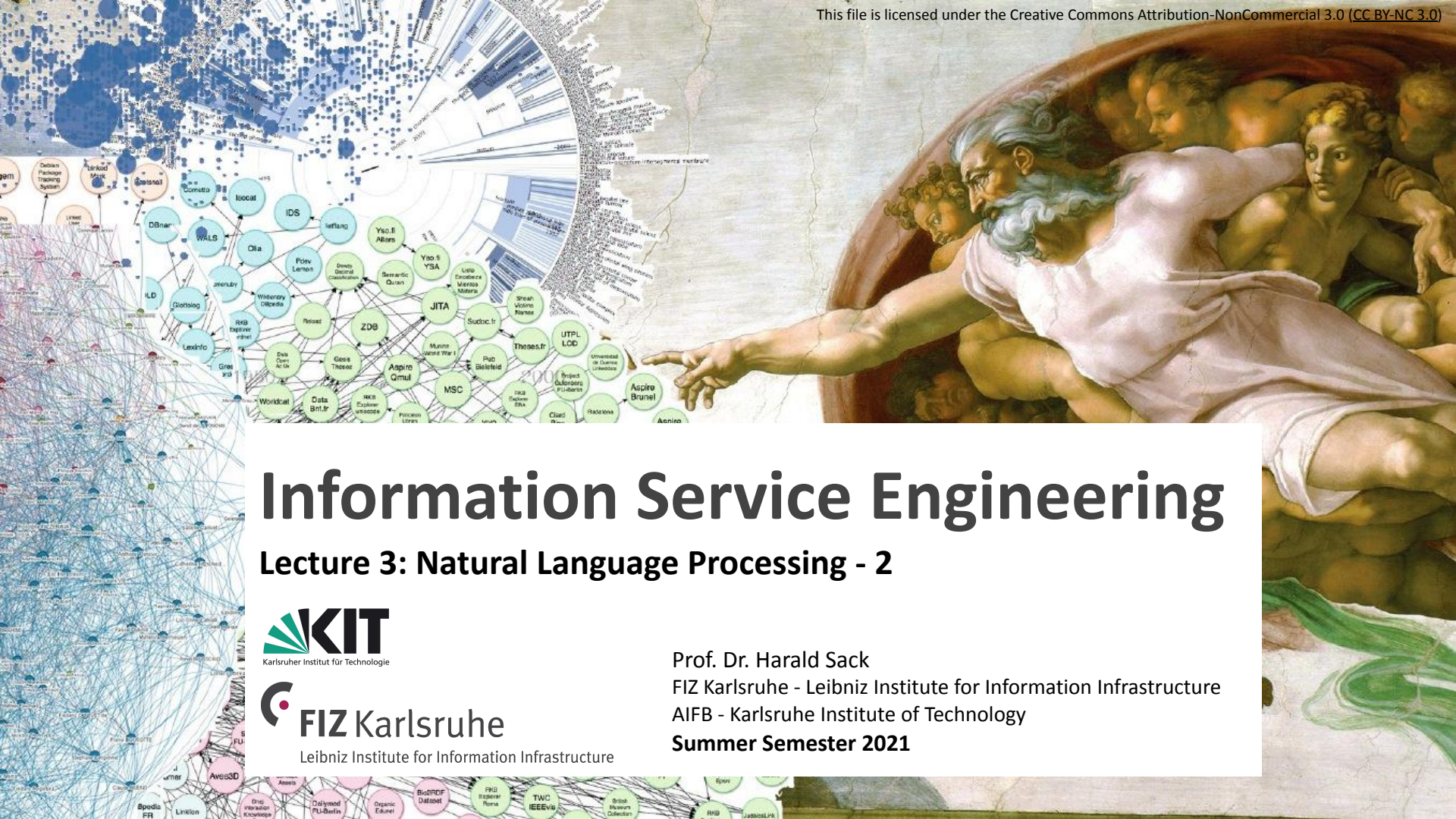
This file is licensed under the Creative Commons Attribution-NonCommercial 3.0 (CC BY-NC 3.0)



Information Service Engineering

Lecture 3: Natural Language Processing - 2

Karlsruher Institut für Technologie

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Prof. Dr. Harald Sack
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
AIFB - Karlsruhe Institute of Technology

Summer Semester 2021

Information Service Engineering

Last Lecture: Natural Language Processing (1)

2.0 What is Natural Language Processing?
2.1 NLP and Basic Linguistic Knowledge
2.2 Morphology
2.3 NLP Applications
2.4 NLP Techniques
2.5 NLP Challenges
2.6 Evaluation, Precision and Recall
2.7 Regular Expressions
2.8 Finite State Automata
2.9 Tokenization
2.10 Language Model and N-Grams
2.11 Part-of-Speech Tagging
2.12 Word Embeddings

- Phonology
- Morphology
- Morphemes
- Free and Bound Morphemes
- Affixes, Prefixes, Suffixes
- Derivation, Compounding, and Inflection
- Stemming and Lemmatization
- NLP Applications
- NLP Techniques

Information Service Engineering

Lecture 3: Natural Language Processing (2)

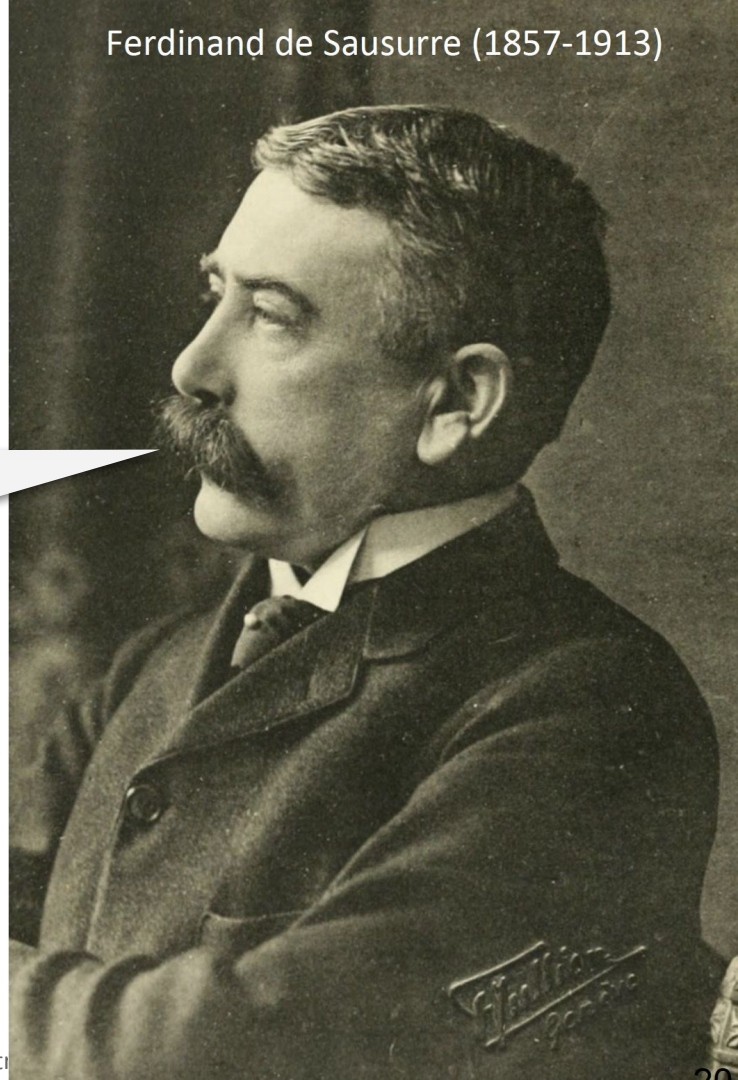
2.0 What is Natural Language Processing?
2.1 NLP and Basic Linguistic Knowledge
2.2 Morphology
2.3 NLP Applications
2.4 NLP Techniques
2.5 NLP Challenges
2.6 Evaluation, Precision and Recall
2.7 Regular Expressions
2.8 Finite State Automata
2.9 Tokenization
2.10 Language Model and N-Grams
2.11 Part-of-Speech Tagging
2.12 Word Embeddings

Why Natural Language is so difficult

I am a Linguist.

I love language more than most people.

1. Paraphrasing
2. Ambiguity



Paraphrasing

- A **paraphrase** is a restatement of the meaning of a text or passage by using other words.
- From Greek παράφρασις, meaning *"additional manner of expression"*
- Examples:
  - Google *bought* YouTube. ⇔ Google *acquired* YouTube.
  - When will my book arrive? ⇔ When will I receive my book?
  - Pat said, "*I like football.*" ⇔ Pat said that *he liked football*.
  - Pat likes Chris, because *she* is smart. ⇔ Pat likes Chris, because *Chris* is smart.

Rahul Bhagat, Eduard Hovy, What Is a Paraphrase?, in Computational Linguistics, Volume 39, Number 3, 2013, doi:10.1162/COLI a 00166

Ambiguity

- One word/sentence can have **different meanings** (in the language to which it belongs to).
- Examples:
  - *“plant”* →
    … plant …
    … workers at the plant …
    … plant a garden …
    … plant meltdown …
    … graze … plant …
    … house plant …
    … CIA plant …
    … plant firmly on the ground …
  - “*The door is open!*”
  - “*We saw her duck.*”
  - “*Kids make nutritious snacks.*”
  - “*It knows you like your mother.*”

Phonological Ambiguities

- Words which **sound the same** but have **different meanings**
  - e.g., *weekend* vs. *weak end.*

Communication tip:

Phonological ambiguities or **Give peas a chance!**

One of my favourite ways to have fun with communication are phonological ambiguities.
Phonological ambiguities are two or more words which sound the same and have different meanings.

Language can contain ambiguities - and more than one way to compose a set of sounds into words.

So listen to yourself: It is always good to notice a spoken sentence often contains many words which are (sometimes not) intended to be heard.

English examples:

- there – their
- here - hear
- plane – plain
- Hamburger (Citizens of Hamburg) – hamburger (burger, food)
- sea - see
- Friday - fry day
- weekend - weak end
- ice cream - I scream.
- new direction - nude erection
- new day - nude, eh?
- I don't know! - I don't - no!
- but – butt
- Wait - Weight
- psychotherapist - psycho the rapist
- You're unconscious now… - Your unconscious now…
- Your students… - You're students…
- Two - too - to

German examples:

- Du hast Gewehre. (You have got guns.) - Du hasst Gewehre. (You hate guns.)
- Lehrer (teacher) – leerer (emptier)

http://worldsgreatestsmile.com/html/phonological_ambiguity.html

Lexical Ambiguities

- **Polysemy** (lexical ambiguity):
  - One word of a **specific syntactic category** can have several meanings, which is in this context called a **lexical sense**,
  - e.g. “*jaguar*”.

- **Homonymy**:
  - Different words that are spelled and pronounced the same way,
  - e.g. *a “book”* vs. *to “book”*,
  - *Time flies like an arrow.*
    Fruit flies like banana.

Syntactic Ambiguities

- A situation where a sentence may be interpreted in more than one way due to **ambiguous sentence structure**.
- Also called *amphiboly* or *amphibology.*
- Example:
  - *He saw the man on the rooftop with a binocular.*
    - ***He saw** the man on the rooftop **with a binocular.***
    - *He saw **the man** on the rooftop **with a binocular**.*

https://upload.wikimedia.org/wikipedia/commons/1/16/Britain%27s_Home_Front_1939_-_1945-_Roof_Spotters_HU86169.jpg

Syntactic Ambiguities

One morning I shot an elephant in my pyjamas. How he got into my pyjamas, I'll never know.

- *Groucho Marx, Animal Crackers (1930)*

https://commons.wikimedia.org/wiki/File:Groucho_Marx.jpg

Semantic Ambiguities

- A situation where a sentence contains **one or more words with different meanings**, and the **significance** of the sentence **changes dramatically** depending on which meaning is intended.
- Example:
  - *"The astronomer loves the **star**."*

White Dwarfs in Globular Cluster NGC 6397, NASA Hubble

Publicity photo of Theda Bara for the American film Cleopatra (1917)

Referential Ambiguities

- In the analysis of coherent sequences of sentences (discourse analysis), subclauses or subsequent clauses might refer to different entities of the first sentence, i.e. **referential ambiguities**.
- Example:
  - *"Alice understands that you like your mother, but she …"*

  - *Does "she" refer to Alice or to your mother?*

Natural Language is always highly ambiguous

- Meaning is **context sensitive**:
  - Depends on the people present e.g. *"How far is it?" (miles, km?)*
  - Depends on the social context: *"That was expensive!".*
  - Depends on the location, e.g. *"Play <song> upstairs".*
  - Depends on the time of day, e.g. *"Let's go eat".*
  - Depends on prior sentences, e.g. *"The third one".*
- It is even more difficult to detect and correctly interpret **slang, jargon, humor, and sarcasm**.
- Also spelling mistakes, grammar mistakes, and (newly created) abbreviations have to be resolved.

Why NLP is really hard...

- The famous **Winograd Schema Challenge** showcases the necessity to combine linguistic and **common-sense/world knowledge** to really understand the semantics of natural language.
- ***"The trophy doesn't fit into the brown suitcase because it's too [small/large]."***
- Answer: ***small*** = suitcase, ***large*** = trophy.

Requires:

1. Anaphora resolution
   (resolution of "***it***" to the correct object depending on the adjective)
2. The knowledge that the smaller object can fit into the larger but not vice versa.
3. The knowledge that a suitcase cannot fit into a trophy.

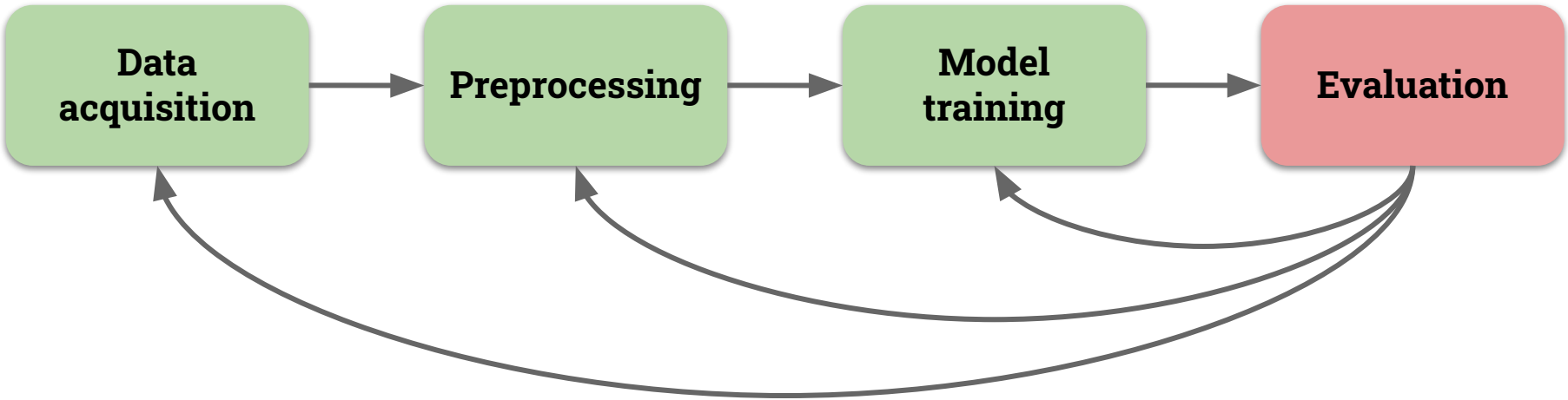
E. Davis, L. Morgenstern, C. Ortiz, *The Winograd Schema Challenge* (2017)

Information Service Engineering

Lecture 3: Natural Language Processing (2)

2.0 What is Natural Language Processing?
2.1 NLP and Basic Linguistic Knowledge
2.2 Morphology
2.3 NLP Applications
2.4 NLP Techniques
2.5 NLP Challenges
2.6 Evaluation, Precision and Recall
2.7 Regular Expressions
2.8 Finite State Automata
2.9 Tokenization
2.10 Language Model and N-Grams
2.11 Part-of-Speech Tagging
2.12 Word Embeddings

NLP in Real-World Applications



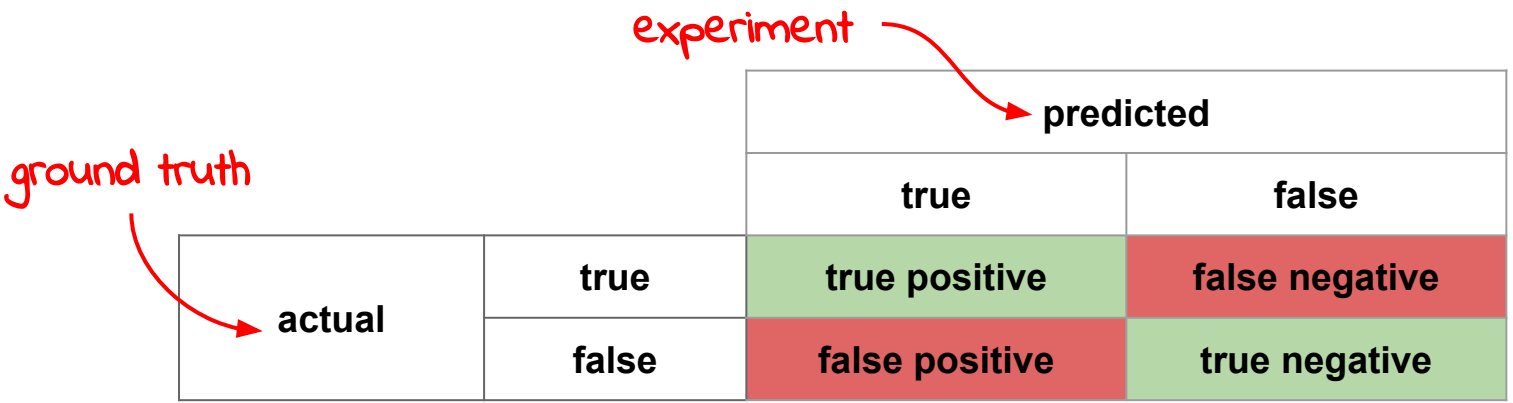
- Building NLP systems is an **iterative cycle**.
- Usually, it is composed of **Human & Machine Learning**.

Evaluation

- How to **objectively measure the quality** of a (classification) experiment?
  - Compare your achieved results with a **ground truth** (**gold standard**).
- How to **achieve a ground truth**?
  - Often this means to invest **manual effort**…
- **How to compare** achieved results with a ground truth?
  - Correctness **Precision**
  - Completeness **Recall**
  - Correctness & Completeness **F-Measure**

Confusion Matrix

- Contains information about **actual and predicted classifications** done by a classification system.
- A table with two rows and two columns that reports the number of
  - **false positives**, **false negatives**, **true positives**, and **true negatives**.

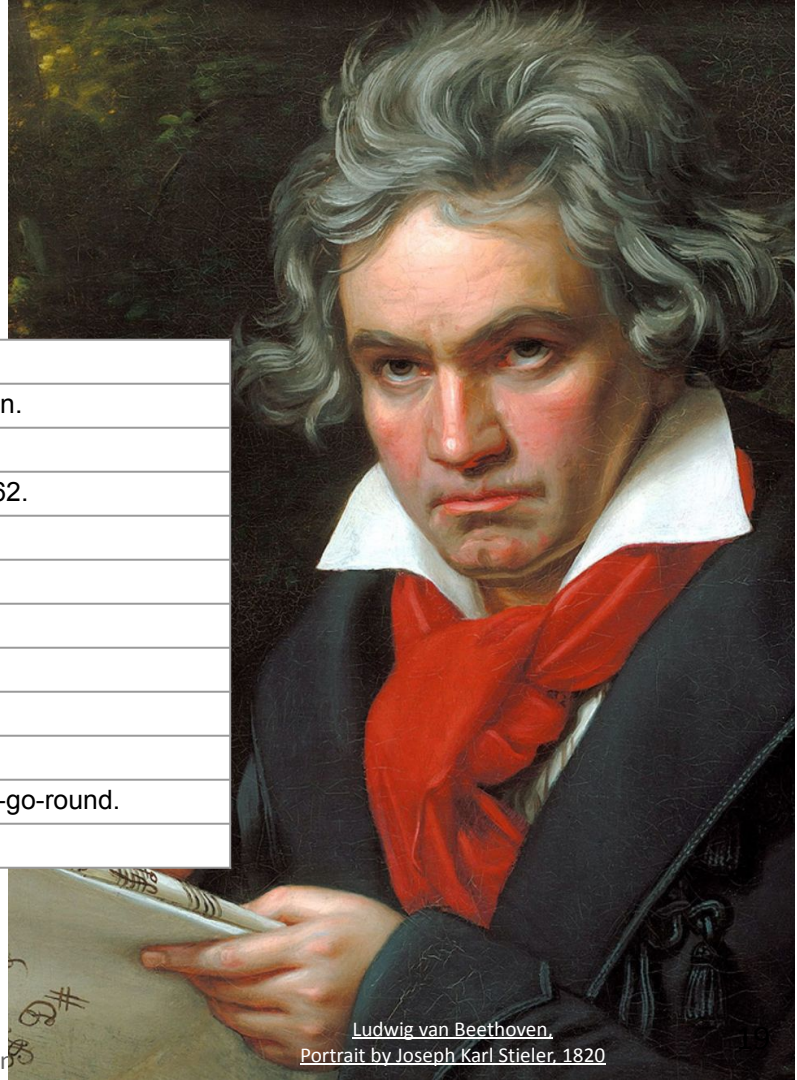


Experiment

- Let's consider the following text corpus:
  BEETHOVENCORPUS

1	The Andante favori is a work for piano solo by Ludwig van Beethoven.
2	The other great passion of the young Mirabehn was the music of van Beethoven.
3	L.V. Beethoven spent the better part of his life in Vienna.
4	Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.
5	Among the few composers writing for the orphica was Ludvig von Beethoven.
6	Betthoven, too, used this key extensively in his second piano concerto.
7	Naue went to Vienna to study briefly with von Beethoven.
8	Bonn is the birthplace of Ludwig van Beethoven (born 1770).
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.
10	Camper van Beethoven were inactive between late 1990 and 1999.
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.
12	Beetehoven hit theaters in april 1992.

C. Barrière, Natural Language Processing in a Semantic Web Context, Springer, 2016, p.11
http://bit.ly/Beethovencorpus



Ludwig van Beethoven,
Portrait by Joseph Karl Stieler, 1820

Experiment

1	The Andante favori is a work for piano solo by Ludwig van Beethoven.
2	The other great passion of the young Mirabehn was the music of van Beethoven.
3	L.V. Beethoven spent the better part of his life in Vienna.
4	Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.
5	Among the few composers writing for the orphica was Ludvig von Beethoven
6	Betthoven, too, used this key extensively in his second piano concerto.
7	Naue went to Vienna to study briefly with von Beethoven.
8	Bonn is the birthplace of Ludwig van Beethoven (born 1770).
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.
10	Camper van Beethoven were inactive between late 1990 and 1999.
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.
12	Beetehoven hit theaters in april 1992.

- **Task**: Identify sentences that refer to **Ludwig van Beethoven**

Experiment

1	The Andante favori is a work for piano solo by Ludwig van Beethoven.	**Actual positive**
2	The other great passion of the young Mirabehn was the music of van Beethoven.	
3	L.V. Beethoven spent the better part of his life in Vienna.	
4	Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.	
5	Among the few composers writing for the orphica was Ludvig von Beethoven	
6	Betthoven, too, used this key extensively in his second piano concerto.	
7	Naue went to Vienna to study briefly with von Beethoven.	
8	Bonn is the birthplace of Ludwig van Beethoven (born 1770).	
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.	**Actual negative**
10	Camper van Beethoven were inactive between late 1990 and 1999.	
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.	
12	Beetehoven hit theaters in april 1992.	

- **Task**: Identify sentences that refer to **Ludwig van Beethoven**

Experiment

1	The Andante favori is a work for piano solo by Ludwig van Beethoven.	**Actual positive**
2	The other great passion of the young Mirabehn was the music of van Beethoven.	
3	L.V. Beethoven spent the better part of his life in Vienna.	
4	Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.	
5	Among the few composers writing for the orphica was Ludvig von Beethoven	
6	Betthoven, too, used this key extensively in his second piano concerto.	
7	Naue went to Vienna to study briefly with von Beethoven.	
8	Bonn is the birthplace of Ludwig van Beethoven (born 1770).	
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.	**Actual negative**
10	Camper van Beethoven were inactive between late 1990 and 1999.	
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.	
12	Beetehoven hit theaters in april 1992.	

- **Task**: Identify sentences that refer to Ludwig van Beethoven.
- **Baseline Algorithm**: *Exact String Match* with full name "***Ludwig van Beethoven***".

Experiment

1	The Andante favori is a work for piano solo by **Ludwig van Beethoven**.	**Actual positive**
2	The other great passion of the young Mirabehn was the music of van Beethoven.	
3	L.V. Beethoven spent the better part of his life in Vienna.	
4	Charles Munch conducted the symphony no. 9 of **Ludwig van Beethoven** in 1962.	
5	Among the few composers writing for the orphica was Ludvig von Beethoven	
6	Betthoven, too, used this key extensively in his second piano concerto.	
7	Naue went to Vienna to study briefly with von Beethoven.	
8	Bonn is the birthplace of **Ludwig van Beethoven** (born 1770).	
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.	**Actual negative**
10	Camper van Beethoven were inactive between late 1990 and 1999.	
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.	
12	Beetehoven hit theaters in april 1992.	

- **Task**: Identify sentences that refer to Ludwig van Beethoven.
- **Baseline Algorithm**: *Exact String Match* with full name "*Ludwig van Beethoven*".
  - Identified **3 lines** (1, 4, 8) as **positive**
  - Identified **9 lines** (2, 3, 5, 6, 7, 9, 10, 11, 12) as **negative**

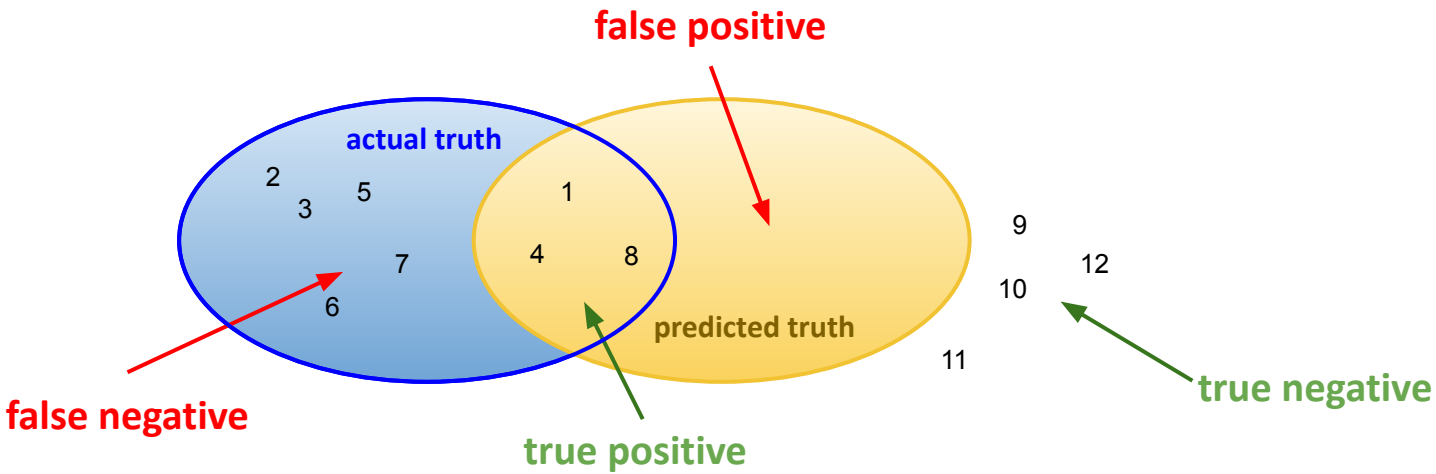
Experiment

- **Baseline Algorithm**: *Exact String Match* with "*Ludwig van Beethoven*".
  - Identified **3 lines** (1, 4, 8) as **positive, all of it are actual positive** (**true positive**)
  - Identified **9 lines** (2, 3, 5, 6, 7, 9, 10, 11, 12) as **negative,**
    - **4 lines** of it (9, 10, 11, 12) are **actual negative (true negative)**
    - **5 lines of it (2,3,5,6,7) are actual positive (false negative)**

		predicted	
		true	**false**
actual	**true**	**3**	**5**
	false	**0**	**4**

Experiment

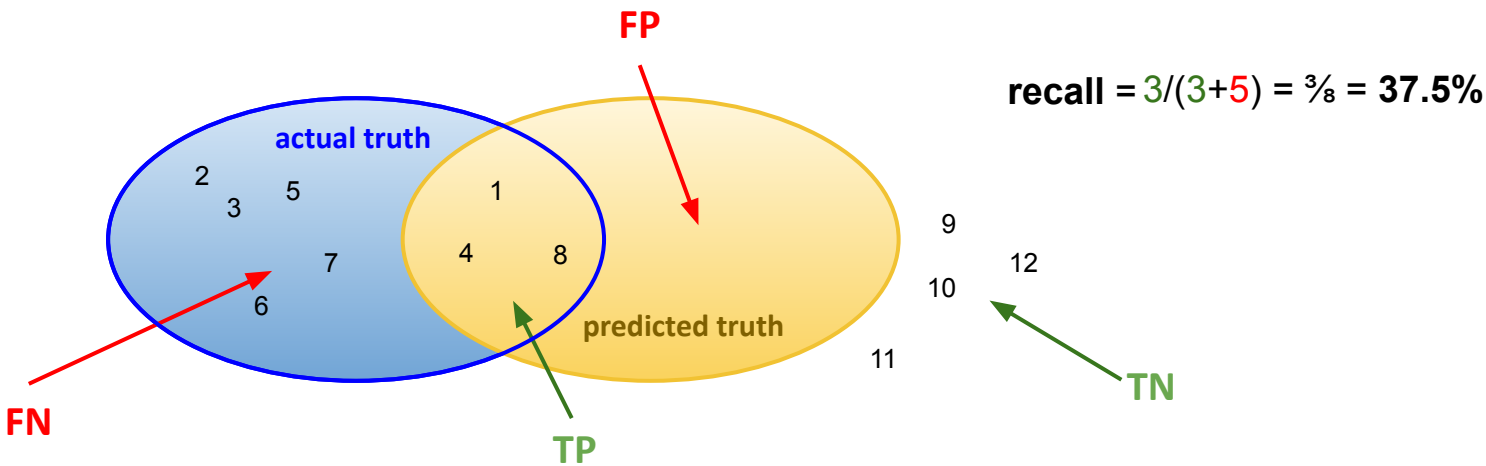
- **Baseline Algorithm**: *Exact String Match* with "*Ludwig van Beethoven*"
  - Identified **3 lines** (1, 4, 8) as **positive, all of it are actual positive** (**true positive, TP**)
  - Identified **9 lines** (2, 3, 5, 6, 7, 9, 10, 11, 12) as **negative,**
    - **4 lines** of it (9, 10, 11, 12) are **actual negative (true negative, TN)**
    - **5 lines of it (2,3,5,6,7) are wrongly identified as negative (false negative, FN)**



Recall

- **Recall** is the **fraction of relevant instances** that are retrieved/predicted.

$$recall = \frac{true\,positive}{true\,positive + false\,negative}$$

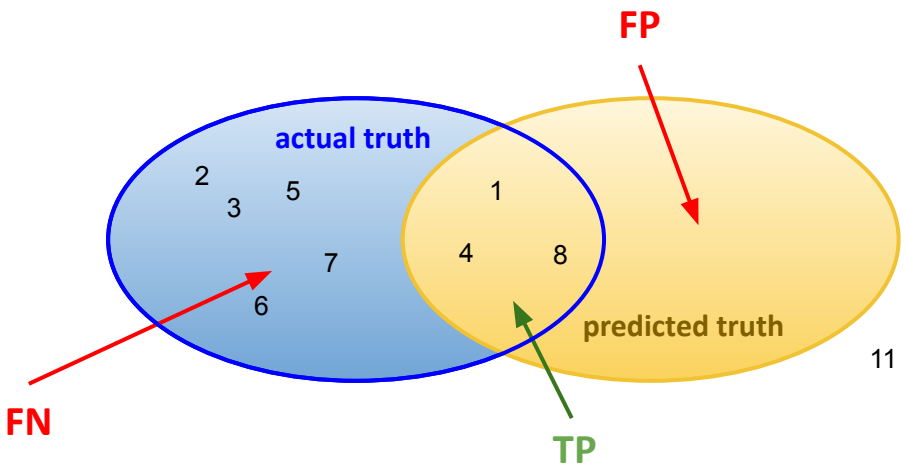


recall = 3/(3+5) = ⅜ = **37.5%**

Precision

- **Precision** is the **fraction of retrieved instances that are relevant.**

$$precision = \frac{true\,positive}{true\,positive + false\,positive}$$

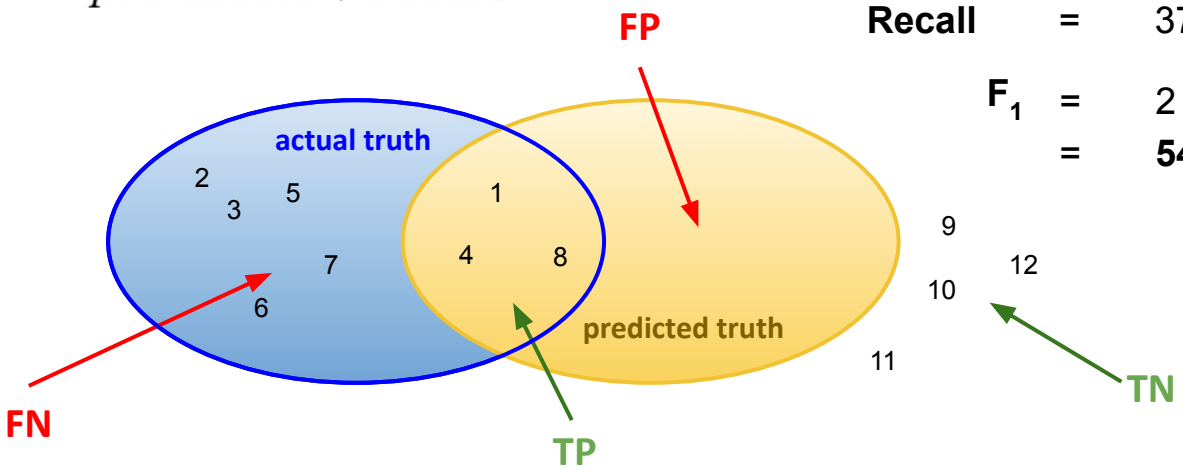


precision =3/(3+0) = 1 = **100%**

F-Measure

- **F-Measure** is a measure that combines precision and recall.
- **F_1-Measure** is the **harmonic mean** of precision and recall.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$



Precision = 100%
Recall = 37.5%

F_1 = 2 * [100 * 37.5 / (100 + 37.5)]
= **54.5 %**

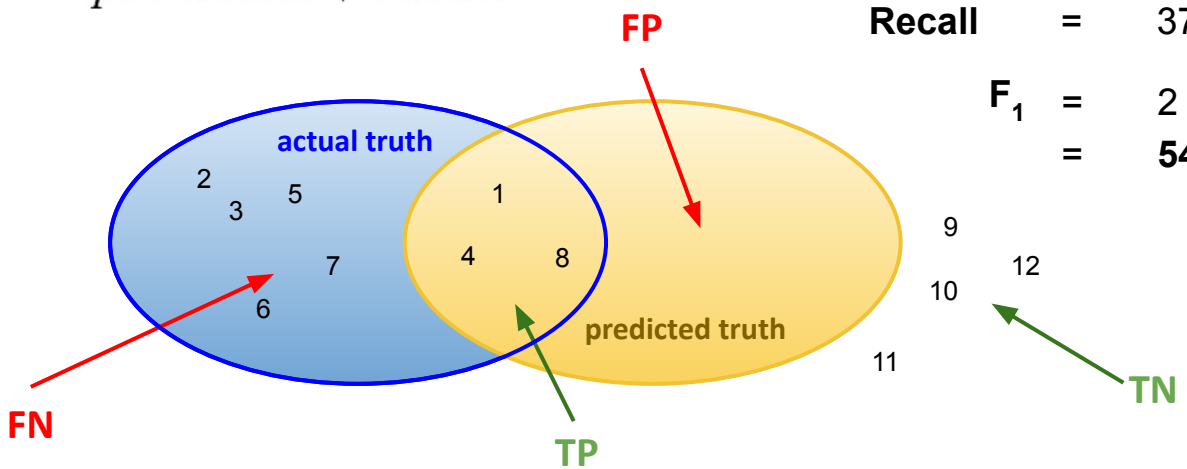
F-Measure

The harmonic mean H of the positive real numbers $x_1, x_2, \ldots, x_n$

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- **F-Measure** is a measure that combines
- **F_1-Measure** is the **harmonic mean** of precision and recall.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$



Precision = 100%
Recall = 37.5%

F_1 = 2 * [100 * 37.5 / (100 + 37.5)]
= **54.5 %**

Experiment

1	The Andante favori is a work for piano solo by Ludwig van **Beethoven**.	**Actual positive**
2	The other great passion of the young Mirabehn was the music of van **Beethoven**.	
3	L.V. **Beethoven** spent the better part of his life in Vienna.	
4	Charles Munch conducted the symphony no. 9 of Ludwig van **Beethoven** in 1962.	
5	Among the few composers writing for the orphica was Ludvig von **Beethoven**	
6	Betthoven, too, used this key extensively in his second piano concerto.	
7	Naue went to Vienna to study briefly with von **Beethoven**.	
8	Bonn is the birthplace of Ludwig van **Beethoven** (born 1770).	
9	Johann van **Beethoven** joined the court, primarily as a singer, in 1764.	**Actual negative**
10	Camper van **Beethoven** were inactive between late 1990 and 1999.	
11	**Beethoven**, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.	
12	Beetehoven hit theaters in april 1992.	

- **Task**: Identify sentences that refer to Ludwig van Beethoven.
- **Another Algorithm**: *Exact String Match* with surname "***Beethoven***".
  - Identified **10 lines** (1,2,3,4,5,7,8,9,10,11) as **positive**
  - Identified **2 lines** (6,12) as **negative**

Ludwig van Beethoven,
Portrait by Joseph Karl Stieler, 1820

Experiment

- **Another Algorithm**: *Exact String Match* with "*Beethoven*".
  - Identified **10 lines** (1,2,3,4,5,7,8,9,10,11) as **positive,**
    - **7 lines** of it (1,2,3,4,5,7,8) **are actual positive** (**true positive**),
    - **3 lines** of it (9,10,11) **are actual negative** (**false positive**)
  - Identified **2 lines** (6, 12) as **negative,**
    - **1 line** of it (12) is **actual negative (true negative)**
    - **1 line** of it (6) is **actual positive (false negative)**

		predicted: true	predicted: false
actual	true	7	1
	false	3	1

- **Precision** = 7/10 = **70%**
- **Recall** = 7/8 = **87.5%**
- **F_1** = 2 (0.7*0.875)/(0.7+0.875)
  = 0.7777 = **77.7%**

Experiment

- **Another Algorithm**: *Exact String Match* with "*Beethoven*".
  - Identified **10 lines** (1,2,3,4,5,7,8,9,10,11) as **positive,**
    - **7 lines** of it (1,2,3,4,5,7,8) **are actual positive** (**true positive**),
    - **3 lines** of it (9,10,11) **are actual negative** (**false positive**)
  - Identified **2 lines** (6, 12) as **negative,**
    - **1 line** of it (12) is **actual negative (true negative)**
    - **1 line** of it (6) is **actual positive (false negative)**

		predicted: true	predicted: false
actual	true	7	1
	false	3	1

- **Precision** = 7/10 = **70%**
- **Recall** = ⅞ = **87.5%**
- F_1 = 2 (0.7*0.875)/(0.7+0.875)
  = 0.7777 = **77.7%**

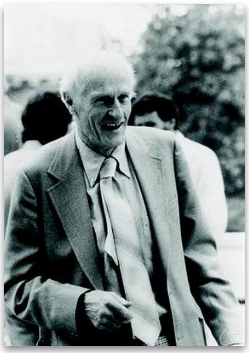
Information Service Engineering

Lecture 3: Natural Language Processing (2)

2.0 What is Natural Language Processing?
2.1 NLP and Basic Linguistic Knowledge
2.2 Morphology
2.3 NLP Applications
2.4 NLP Techniques
2.5 NLP Challenges
2.6 Evaluation, Precision and Recall
2.7 Regular Expressions
2.8 Finite State Automata
2.9 Tokenization
2.10 Language Model and N-Grams
2.11 Part-of-Speech Tagging
2.12 Word Embeddings

Regular Expressions

- **Regular Expressions (RE)** are a formal language to define search patterns.
- RE can be used in **UNIX tools**
  - *grep, sed, awk,...*
- as well as in **programming languages**, as e.g.
  - *Python, Java, .NET*, etc.
- Introduced by Stephen C. Kleene (1951),
  applied for text search first by Ken Thompson (1968).

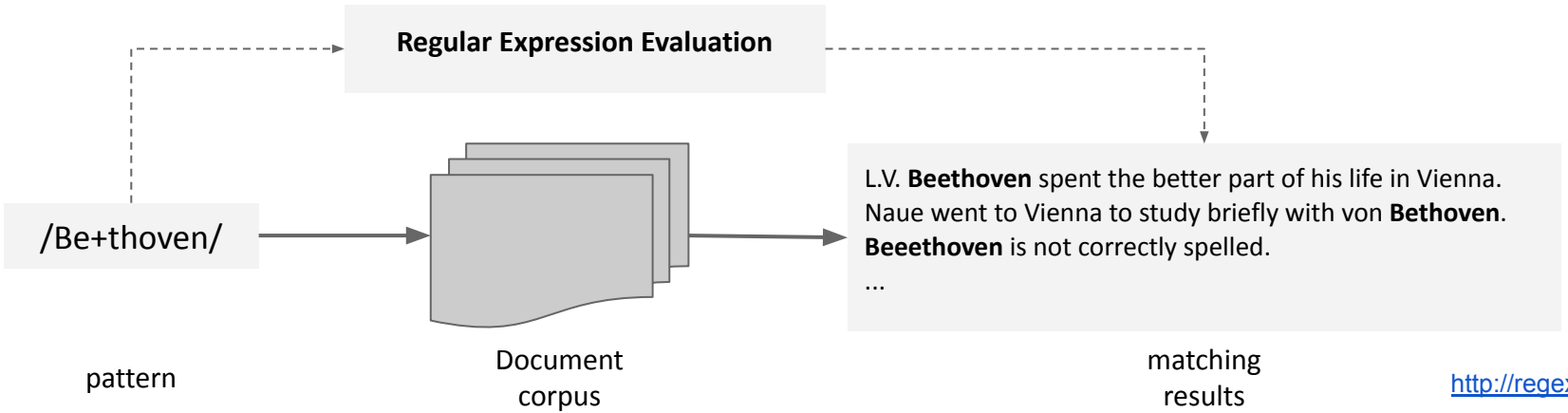


Stephen Cole Kleene
(1909 - 1994)

Kleene, Stephen C. (1951). Shannon, Claude E.; McCarthy, John (eds.). Representation of Events in Nerve Nets and Finite Automata. Automata Studies. Princeton University Press. pp. 3–42.
Ken Thompson (Jun 1968). Programming Techniques: Regular expression search algorithm. Communications of the ACM. 11 (6): 419–422

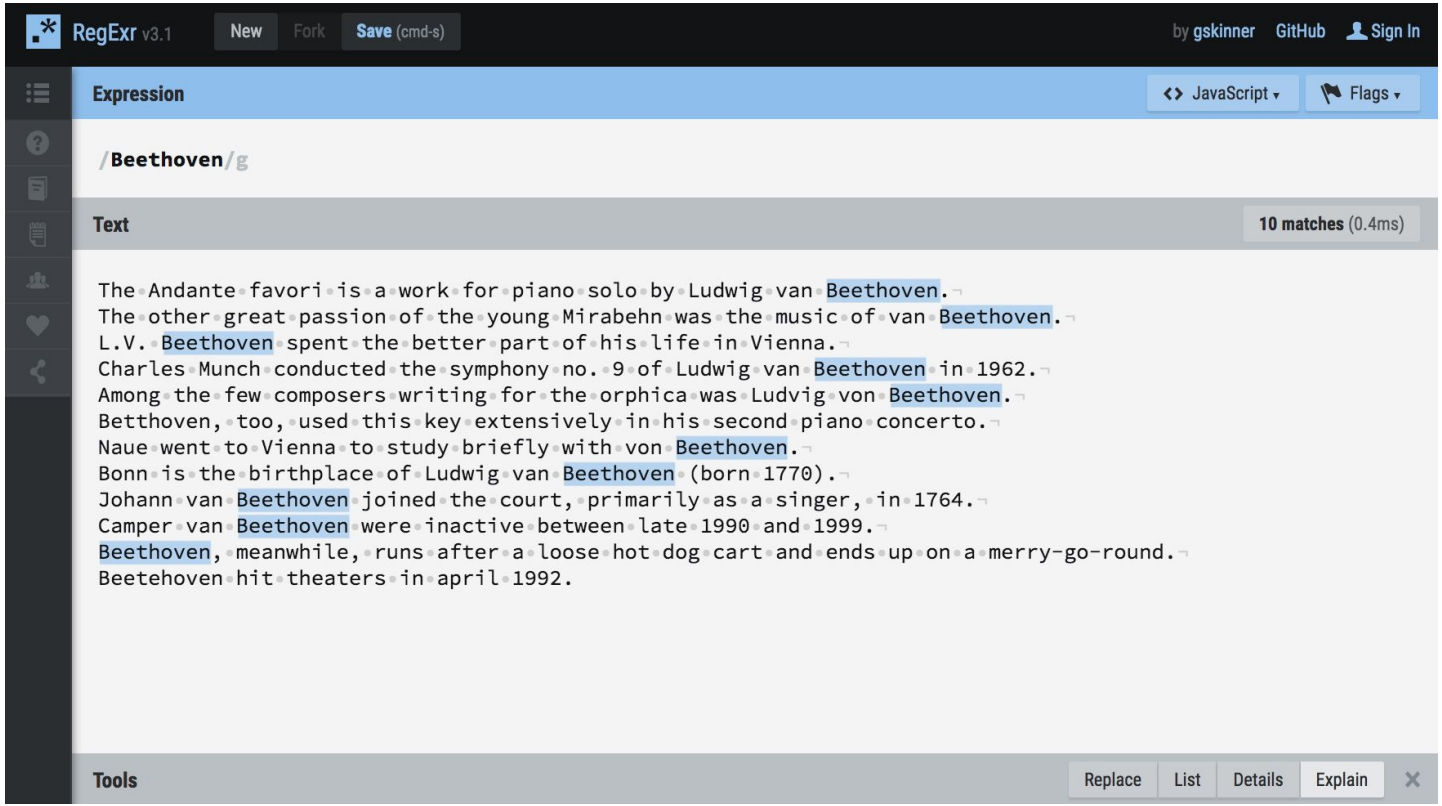
Regular Expressions

- RE are an algebraic notation that specifies simple classes of **strings.**
- A **string** is defined as a sequence of symbols from an alphabet.
- RE search requires a **pattern** that is to be searched and a **corpus** of texts to search through.



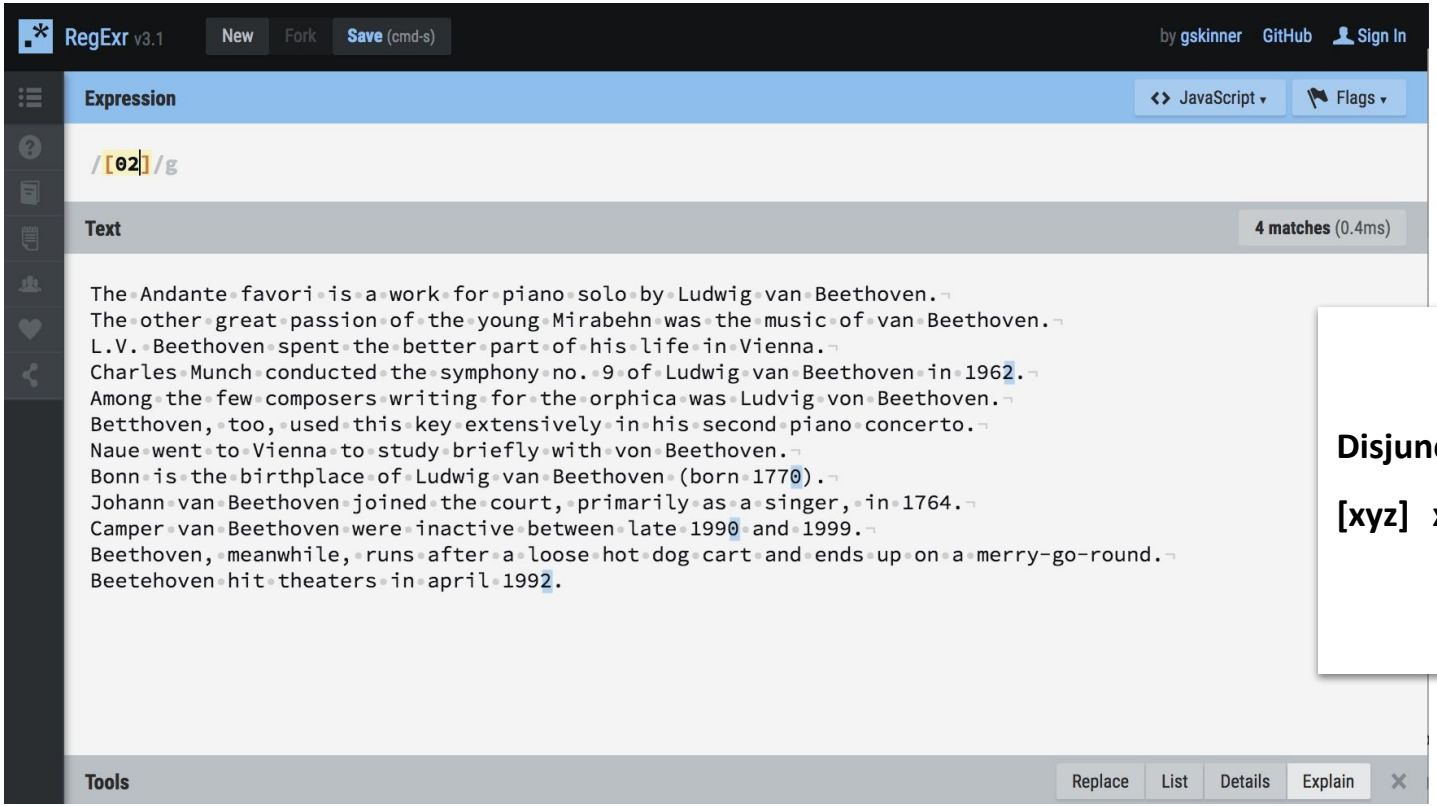
http://regexr.com/

Regular Expressions - Exact String



http://regexr.com/

Regular Expressions - Disjunction

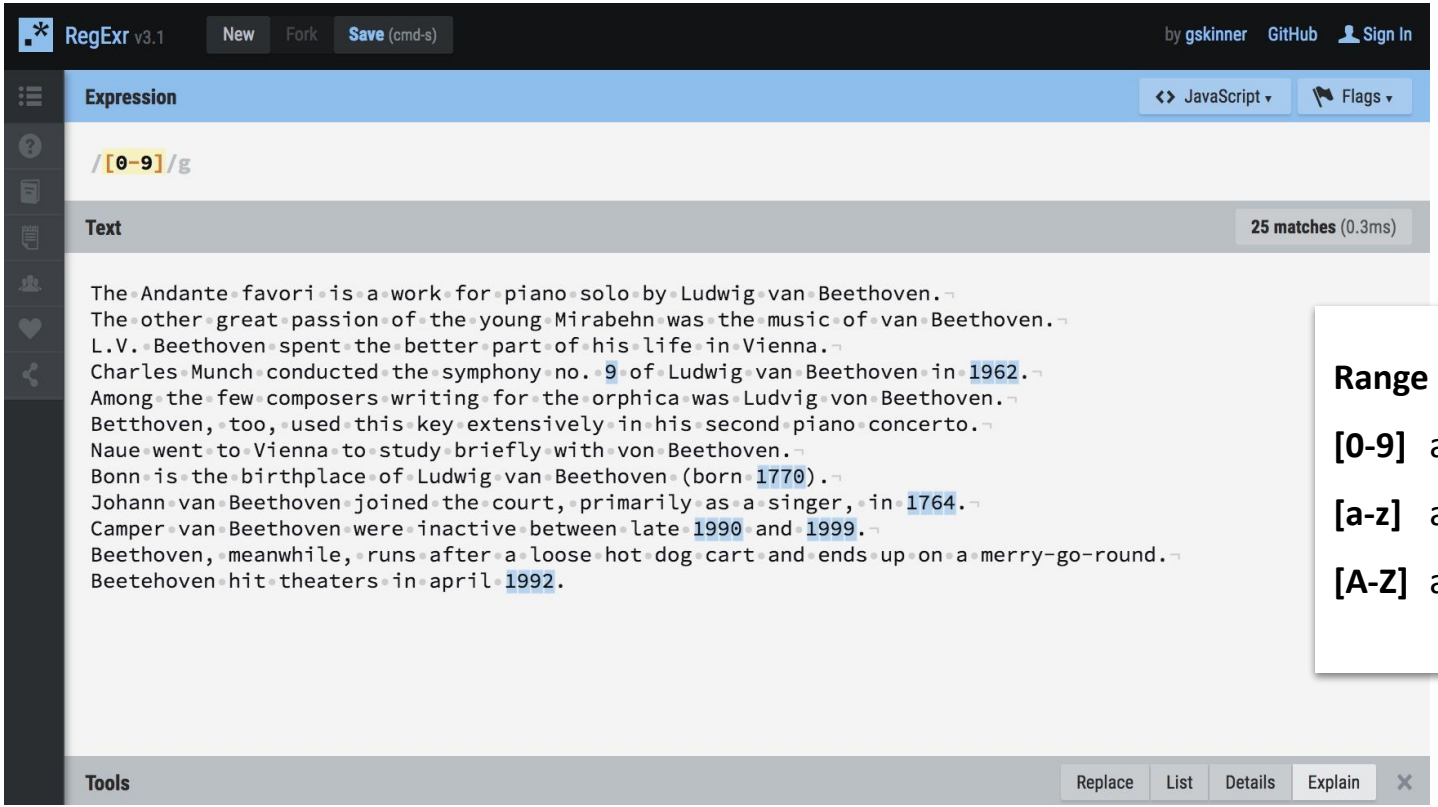


Disjunction

[xyz] **x** or **y** or **z**

http://regexr.com/

Regular Expressions - Ranges



Range Expressions

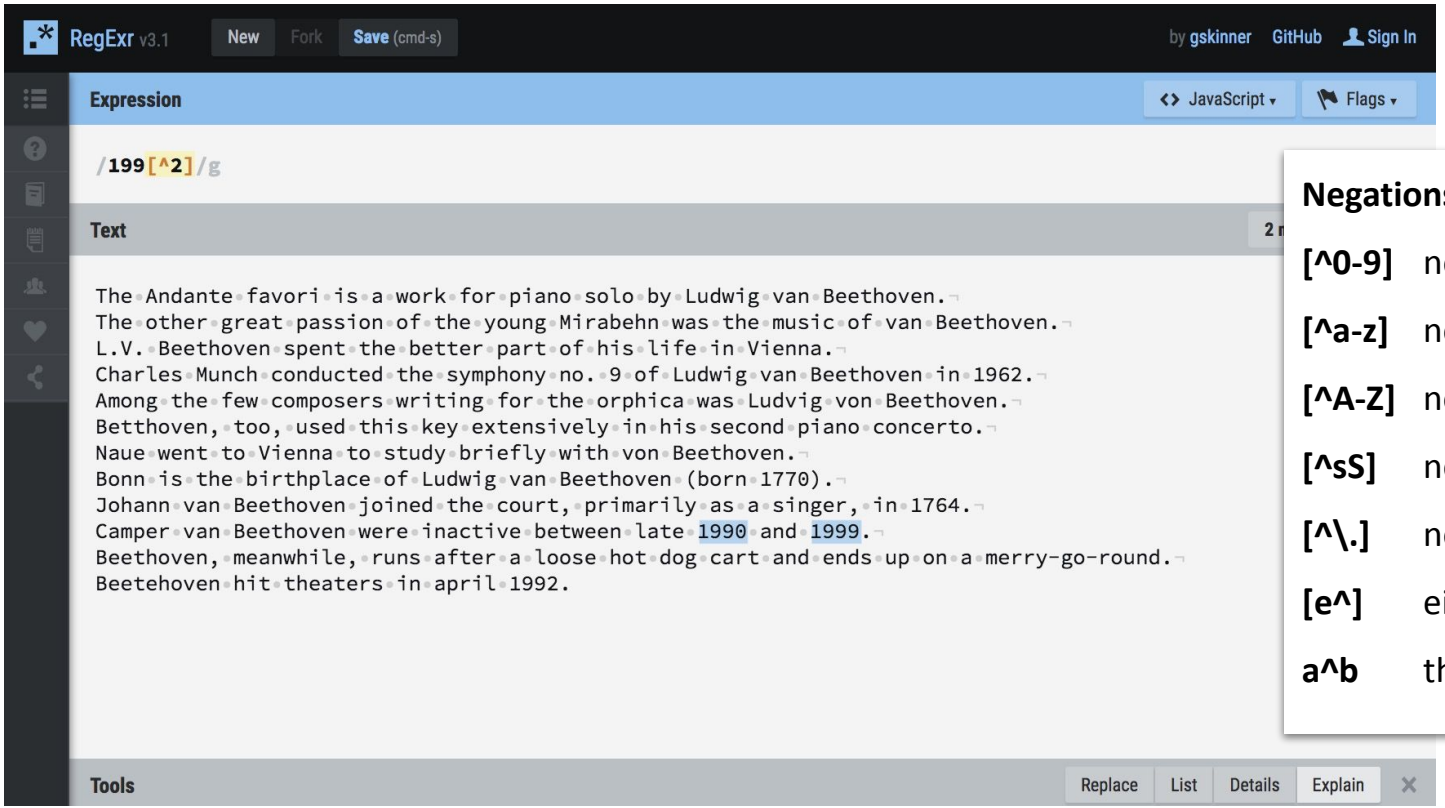
[0-9] any single digit

[a-z] any lower case letter

[A-Z] any upper case letter

http://regexr.com/

Regular Expressions - Negations

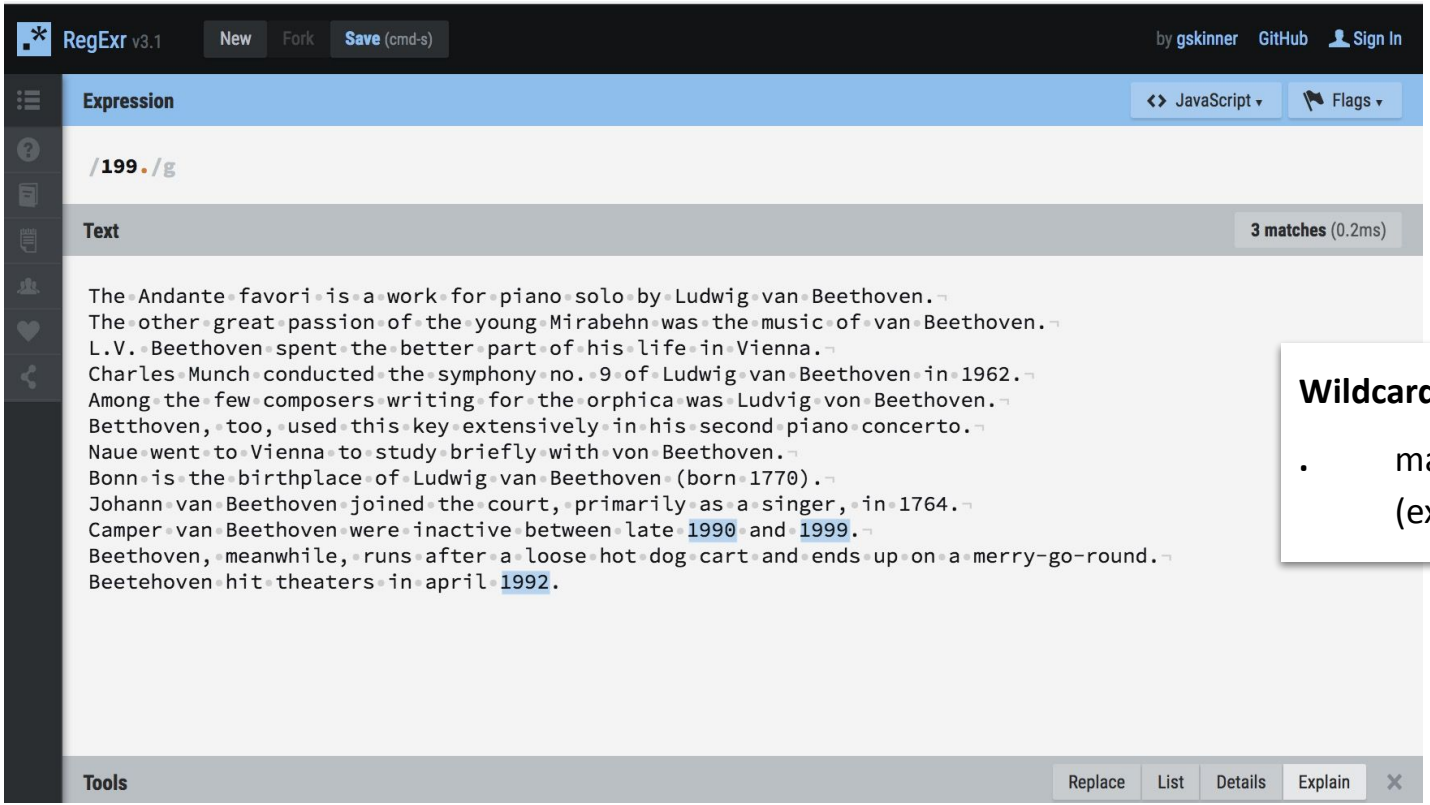


Negations

[^0-9]	not a digit
[^a-z]	not a lower case letter
[^A-Z]	not upper case letter
[^sS]	neither **s** nor **S**
[^\.]	not a period
[e^]	either **e** or **^**
a^b	the pattern **“a^b”**

http://regexr.com/

Regular Expressions - Wildcard

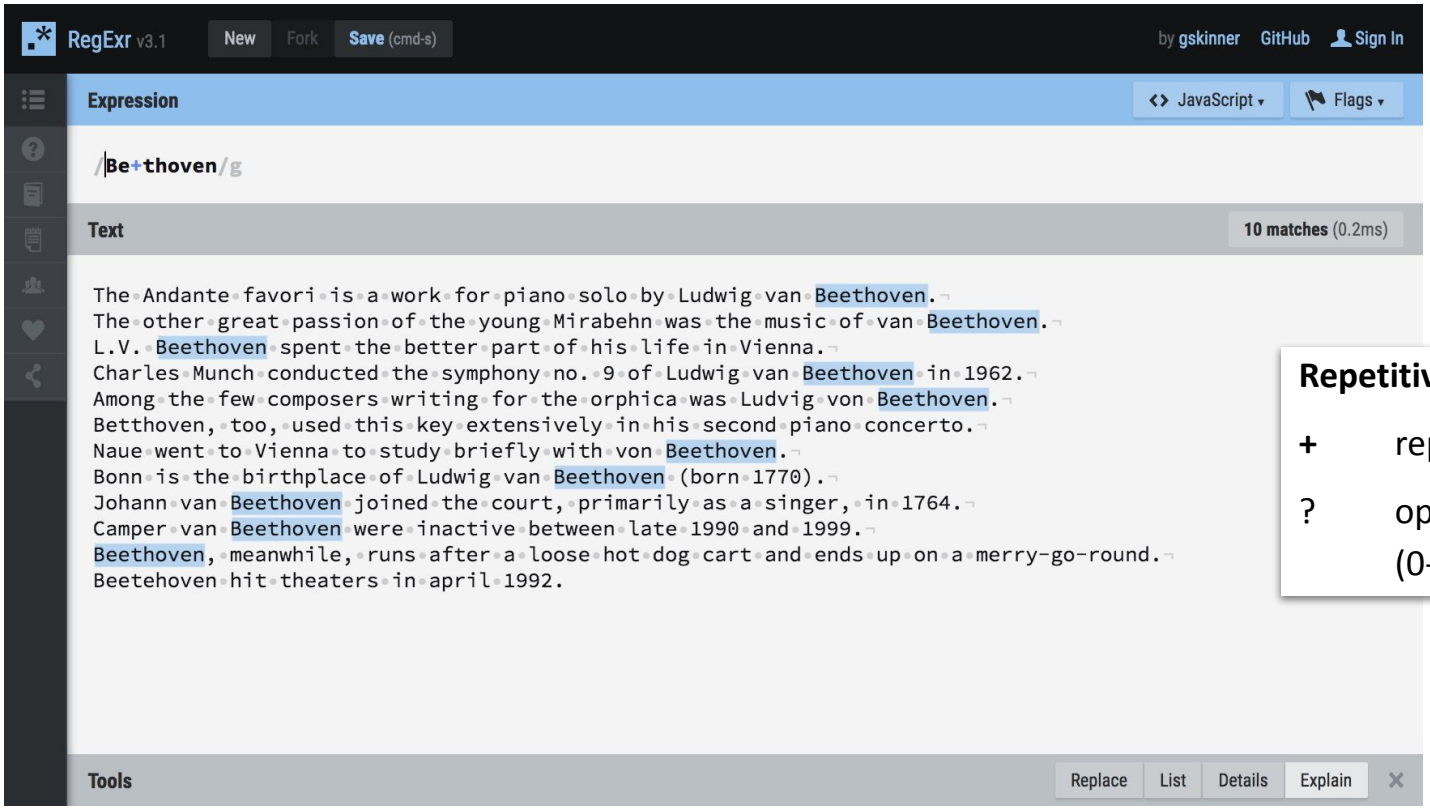


Wildcard

. matches any character (except carriage return)

http://regexr.com/

Regular Expressions - Repetitive Pattern

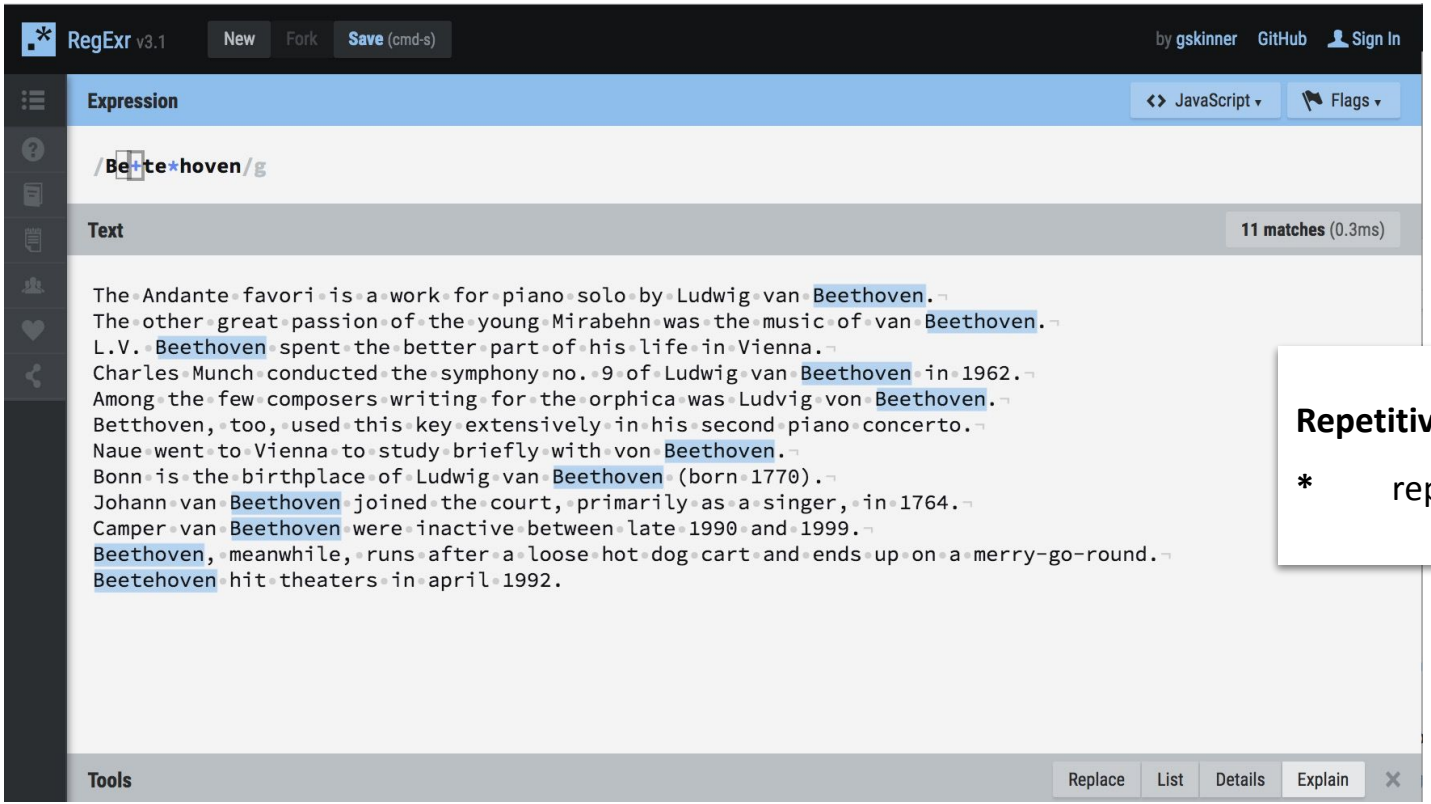


Repetitive Pattern

+	repeat pattern 1-n times
?	optional pattern (0-1 times)

http://regexr.com/

Regular Expressions - Repetitive Pattern

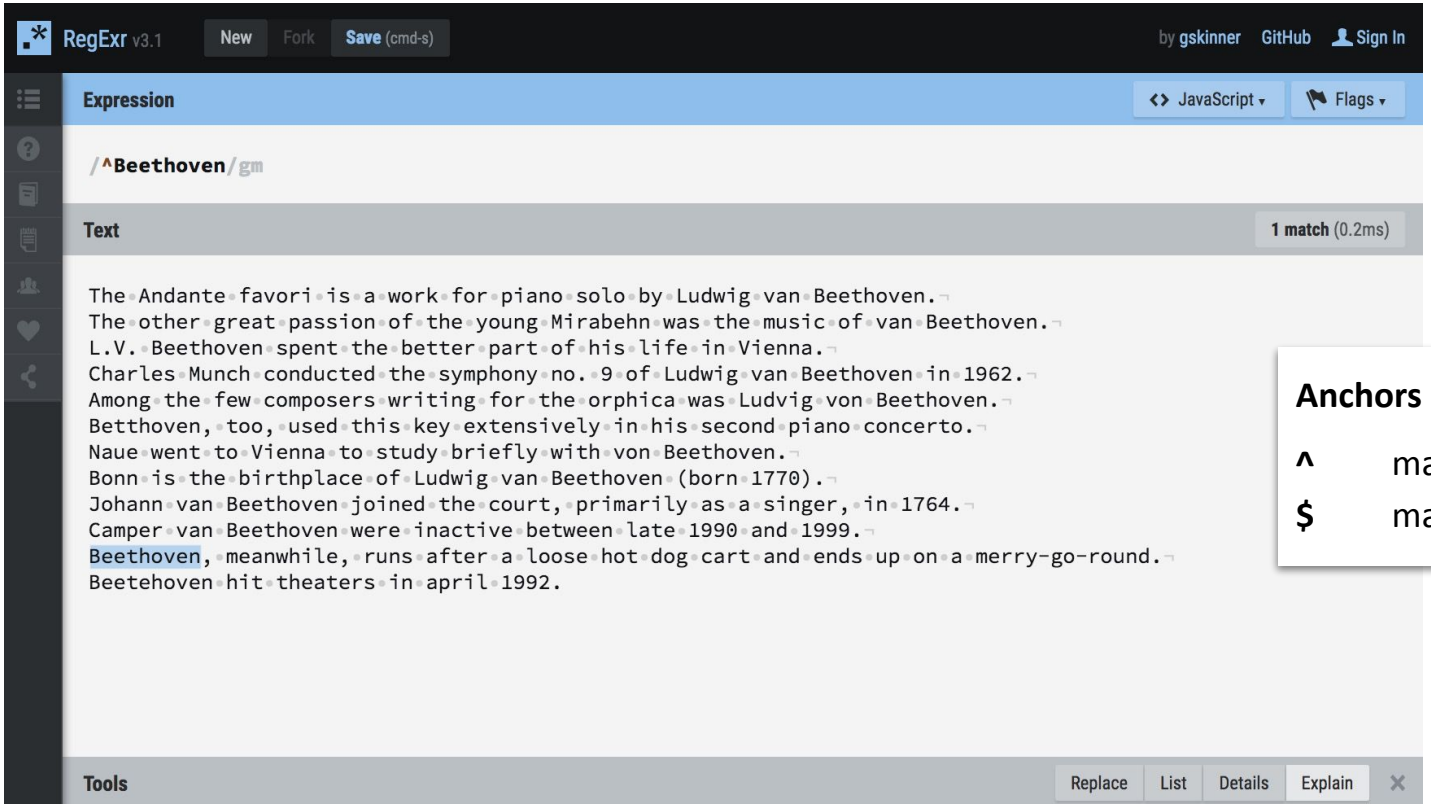


Repetitive Pattern

* repeat pattern 0-n times

http://regexr.com/

Regular Expressions - Anchors

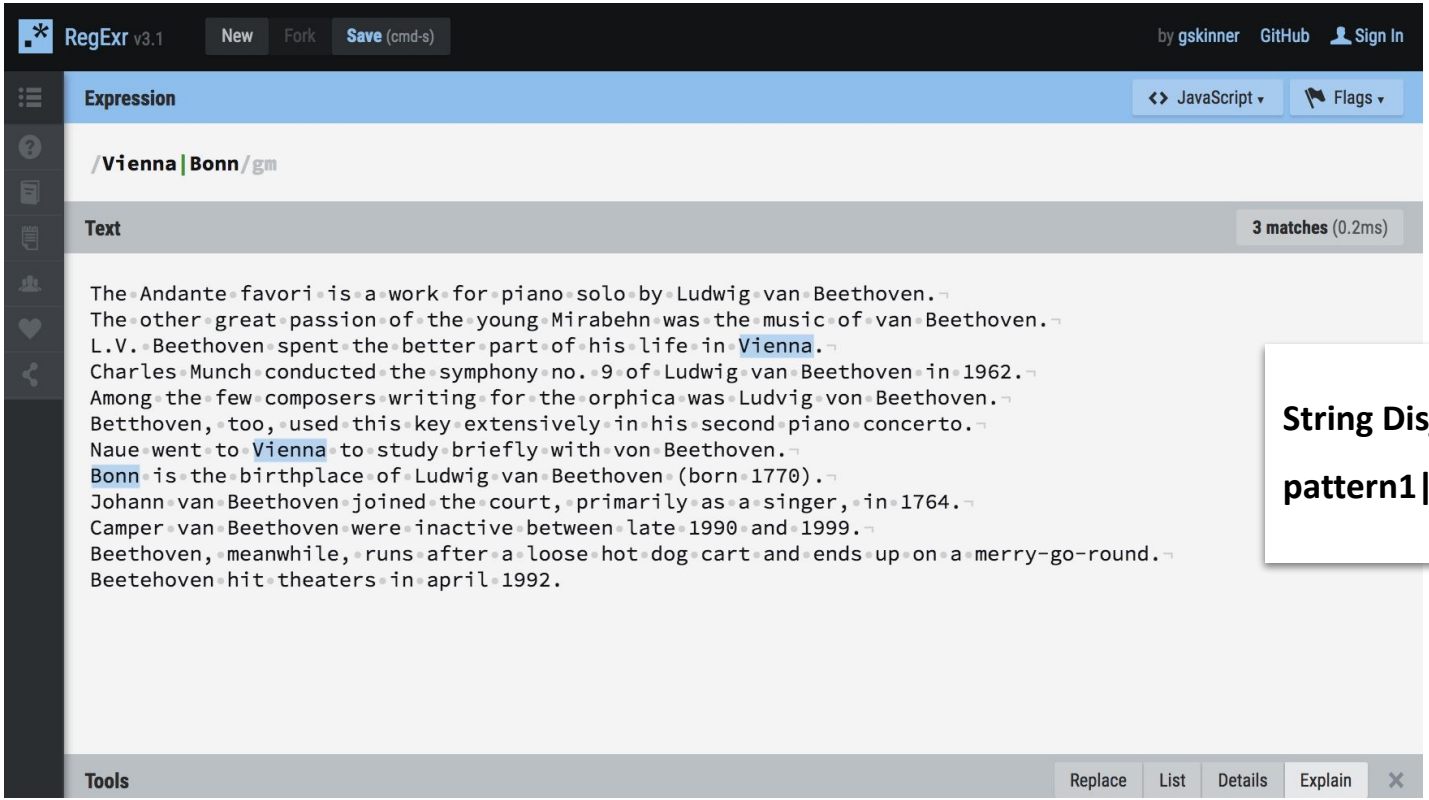


Anchors

^	matches beginning of line
$	matches end of line

http://regexr.com/

Regular Expressions - String Disjunction

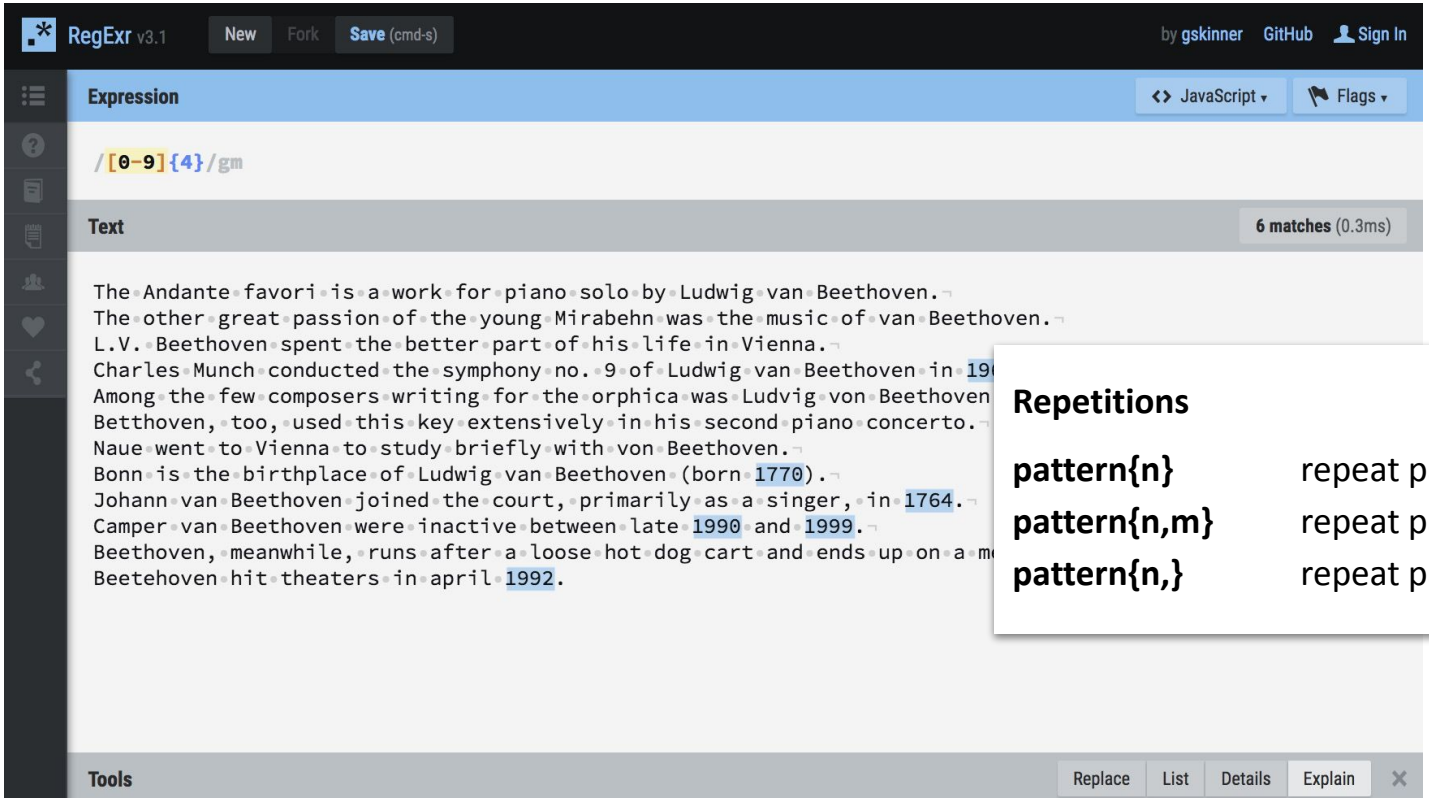


String Disjunction

pattern1|pattern2

http://regexr.com/

Regular Expressions - Repetitions



Repetitions

pattern{n}	repeat pattern n times
pattern{n,m}	repeat pattern n to m times
pattern{n,}	repeat pattern at least n times

http://regexr.com/

Regular Expressions

- Some characters need to be **backslashed**:

RE	Match
`\*`	An asterisk
`\.`	A period
`\?`	A question mark
`\n`	A newline
`\t`	A tab
`\,`	A comma

  - All functional characters that are to be used as '**characters only**' in a pattern must be backslashed.

Regular Expressions

- Advanced operators:

RE	Expansion	Match
\d	[0-9]	Any digit
\D	[^0-9]	Any non digit
\w	[a-zA-Z0-9_]	Any alphanumeric + underscore
\W	[^\w]	Any non-alphanumeric
\s	[\r\t\n\f]	Whitespace
\S	[^\s]	Non-whitespace

Regular Expressions

- Numeric ranges:

RE	Match
`*`	Zero or more occurrences of previous character or expression
`+`	One or more occurrences of previous character or expression
`?`	Exactly zero or one occurrence of previous character or expression
`{n}`	n occurences of previous character or expression
`{n,m}`	From n to m occurrences of previous character or expression
`{n,}`	At least n occurrences of previous character or expression

Detecting Synonyms and Variations

- If we are searching for **all occurrences of an entity in a text**, we have to consider **synonyms** and **variations** of its name:
  - **Real synonyms** e.g. mobile phone -> cell phone, cellular telephone
  - **Quasi synonyms** e.g. mobile phone -> flip phone, mobile
  - **Upper case variations** e.g. cell phone and Cell phone
  - **Orthographic variations** e.g. cell phone and cell-phone
  - **Plural forms** e.g. cell phone and cell phones
  - **Typographic errors** e.g. cellular phone
  - **Related topics** e.g. cellphone video, cellular radio, phone carrier

Synonyms and Variations

1	The Andante favori is a work for piano solo by Ludwig van Beethoven.	**Actual positive**
2	The other great passion of the young Mirabehn was the music of van Beethoven.	
3	L.V. Beethoven spent the better part of his life in Vienna.	
4	Charles Munch conducted the symphony no. 9 of Ludwig van Beethoven in 1962.	
5	Among the few composers writing for the orphica was Ludvig von Beethoven	
6	Betthoven, too, used this key extensively in his second piano concerto.	
7	Naue went to Vienna to study briefly with von Beethoven.	
8	Bonn is the birthplace of Ludwig van Beethoven (born 1770).	
9	Johann van Beethoven joined the court, primarily as a singer, in 1764.	**Actual negative**
10	Camper van Beethoven were inactive between late 1990 and 1999.	
11	Beethoven, meanwhile, runs after a loose hot dog cart and ends up on a merry-go-round.	
12	Beetehoven hit theaters in april 1992.	

- **Task**: Identify sentences that refer to **Ludwig van Beethoven**.
- **Another Algorithm**: *RE Match* with "***Bee*t+hoven***"

Ludwig van Beethoven,
Portrait by Joseph Karl Stieler, 1820

Experiment

- **Another Algorithm**: *RE Match* with "***Bee*t+hoven***"
  - Identified **11 lines** (1,2,3,4,5,6,7,8,9,10,11) as **positive,**
    - **8 lines** of it (1,2,3,4,5,7,8) **are actual positive** (**true positive**),
    - **3 lines** of it (9,10,11) **are actual negative** (**false positive**)
  - Identified **1 line** (12) as **negative,**
    - **1 line** of it (12) is **actual negative (true negative)**

		predicted	
		true	**false**
actual	**true**	**8**	**0**
	false	**3**	**1**

- **Precision** = 8/11 = **72,7%**
- **Recall** = 8/8 = **100%**
- **F_1** = 2 (0.727*1)/(0.727+1)
  = 0.842 = **84.2%**

Experiment

- Can you come up with a **regular expression** that obtains **F_1=100%** for the Beethoven Corpus?
- If so,
  will this be the **“perfect” Beethoven classifier** ?

Ludwig van Beethoven,
Portrait by Joseph Karl Stieler, 1820

Information Service Engineering

Lecture 3: Natural Language Processing (2)

2.0 What is Natural Language Processing?
2.1 NLP and Basic Linguistic Knowledge
2.2 Morphology
2.3 NLP Applications
2.4 NLP Techniques
2.5 NLP Challenges
2.6 Evaluation, Precision and Recall
2.7 Regular Expressions
2.8 Finite State Automata
2.9 Tokenization
2.10 Language Model and N-Grams
2.11 Part-of-Speech Tagging
2.12 Word Embeddings

3. Natural Language Processing - 2

Bibliography

- C. Barrière, *Natural Language Processing in a Semantic Web Context*, Springer, 2016.
- E. Davis, L. Morgenstern, C. Ortiz, *The Winograd Schema Challenge* (2017).
- Rahul Bhagat, Eduard Hovy, *What Is a Paraphrase?*, in Computational Linguistics, Volume 39, Number 3, 2013, doi:10.1162/COLI a 00166.
- D. Jurafsky, J. H. Martin, *Speech and Language Processing*, 2nd ed (draft), 2007,
  - *Section 2, Regular Expressions and Automata*
    (please note that this refers to the 2nd ed.).

3. Natural Language Processing - 2

Syllabus Questions

- What is a paraphrase and why is this difficult for NLP?
- Explain the different forms of ambiguity in natural language.
- How can ambiguity in natural language be solved in general? What additional information is necessary to solve ambiguities?
- What are Winograd Schema Challenges and why are they especially hard?
- How does an arbitrary NLP experiment look like?
- How are recall, precision and f-measure defined?
- Why are recall or precision alone not sufficient measures for the quality of a result?
- What are regular expressions and what can they be used for in NLP?