Natural Language Processing

01: Language and Morphology

1. Linguistic Basics
2. Morphology
3. Evaluation, Precision and Recall
4. Regular Expressions

BORHAN
KNOWLEDGE REASONING

Dr. Imran Ihsan
Ph.D. in Knowledge Engineering
Associate Professor

01-01
Linguistic Basics

01 Language and Morphology

Linguistics

is the scientific study of language and its structure

and involves an analysis of

- language form (phonetics, syntax and grammar),
- language meaning, and
- language in context

Natural Language Processing

is a field of computer science, artificial intelligence, and computational linguistics and

is concerned with the interactions between computers and human (natural) languages and, in particular,

is concerned with programming computers to fruitfully process large natural language corpora.

Specifically, the task to extract meaningful information from natural language input or to produce natural language output.

Phonology

is concerned with the systematic organization of sounds in languages, i.e. the abstract, grammatical characterization of systems of sounds (or signs)

at all levels of language where sound is structured for conveying linguistic meaning.

Phone

any distinct speech sound, regardless of whether the exact sound is critical to the meanings of words

Phoneme

smallest (abstract cognitive) sound unit in a language that is able of conveying a distinct meaning

Example:

"s" and "r" in "sing" and "ring"

"ss" and "ll" in "kiss" and "kill"

Morphology

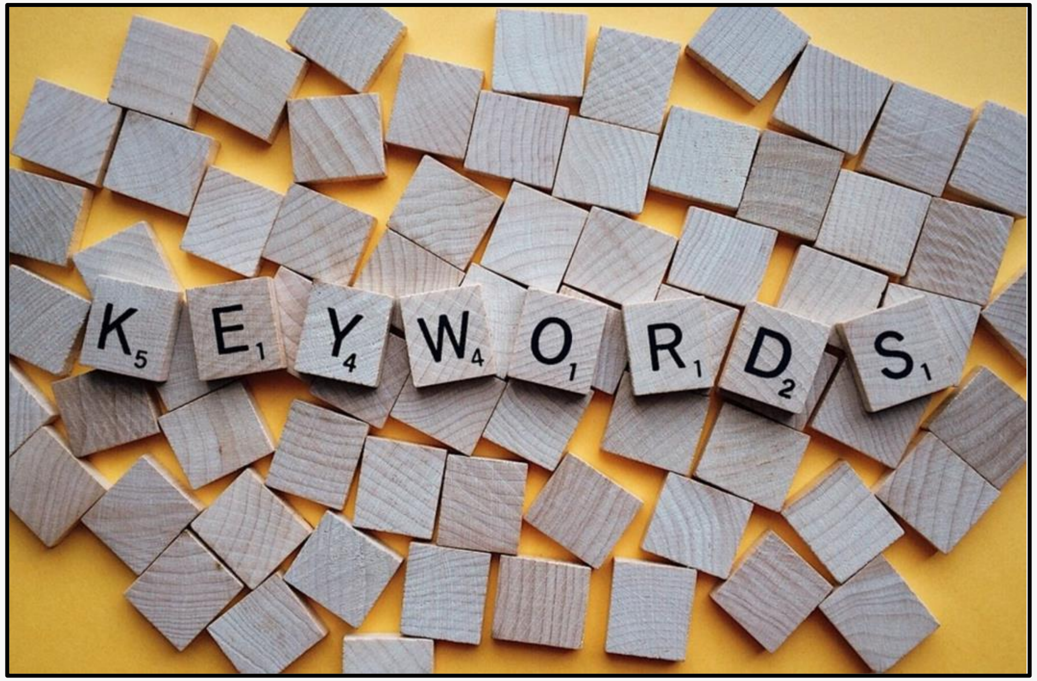
The study of internal structures (formation) of words and how they can be modified

Parsing complex words into their components

⇨ What is a word?

Word and Vocabulary

A word (w_i) is the smallest independent unit of language



"Independent"?

do not depend on other words
can be separated from other units
can change position

Example:

The man looked at the horses

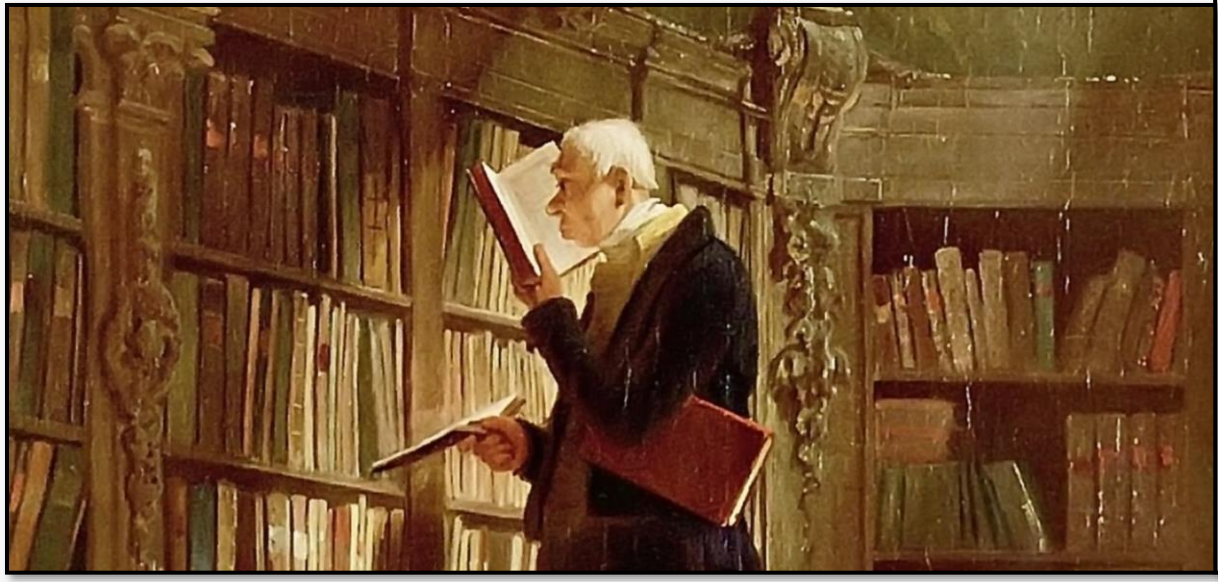
s is the plural marker, dependent on the noun horse to receive meaning
Horses is a word: can occur in other positions or stand on its own

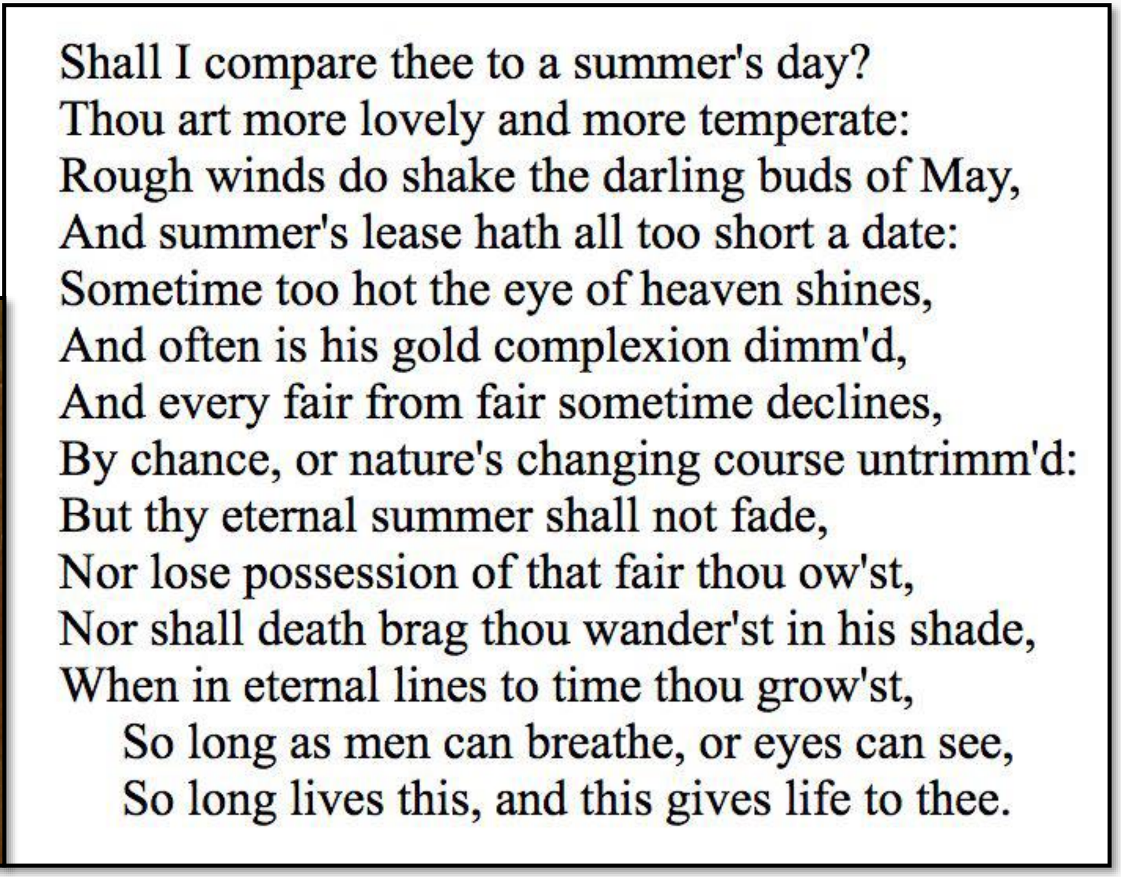
A vocabulary consists of a set of words (w_i)

Text and Language

A text is composed of a sequence of words from a vocabulary

A language is constructed of a set of all possible texts





Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date:
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimm'd,
And every fair from fair sometime declines,
By chance, or nature's changing course untrimm'd:
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

01-02
Morphology

01 Language and Morphology

Morphology

The study of internal structures (formation) of words and how they can be modified.

Parsing complex words into their components (morphemes)

Morphemes

The smallest grammatical unit in a language

i.e. the smallest meaningful unit of a language

We distinguish:

Simple Words: consist of a single morpheme e.g. work, build, run, etc.

Complex Words: have internal structure i.e. exist of 2 or more morphemes

e.g. worker, *affix* –er added to *root* work.

A bound morpheme that is part of a complex word but doesn't belong to any lexical category (i.e., is not a verb, a noun, an adjective)

Core part of a complex word, the part that carries the major component of its meaning

Free vs. Bound Morphemes

Free Morphemes

A simple word, consisting of one morpheme

E.g., house, work, high, chair, wrap

Bound Morphemes

Morphemes that must be attached to another morpheme to receive meaning

E.g., unkindness

Un- and –ness are bound morphemes that require the root kind to receive meaning

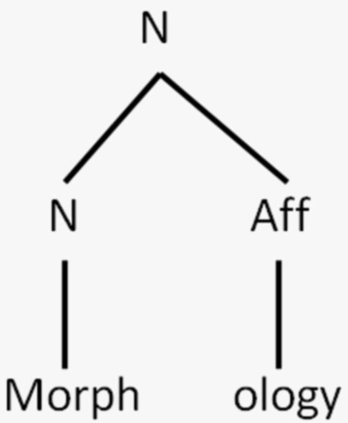
Prefix

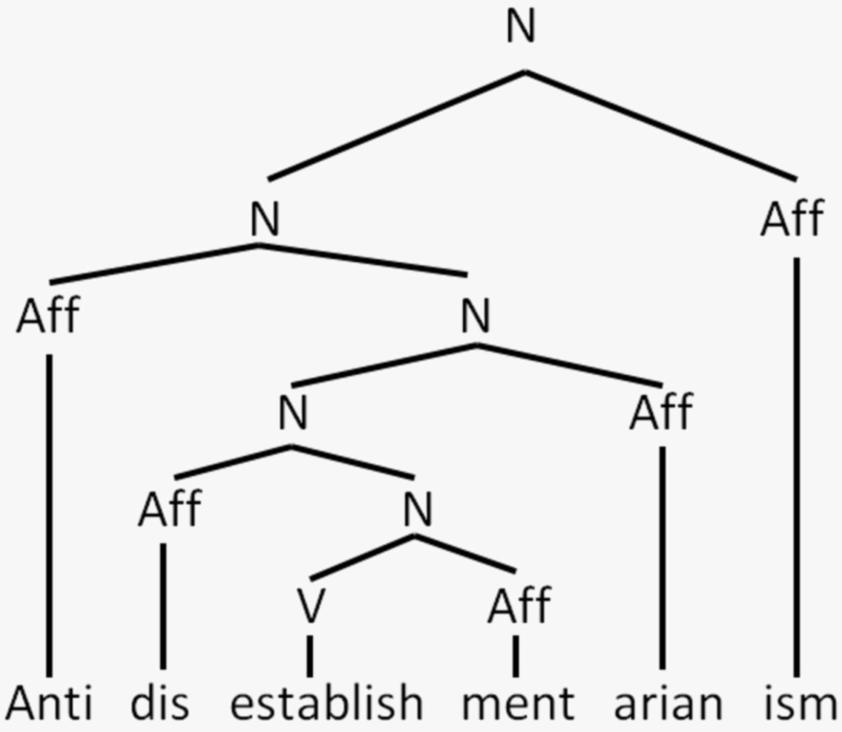
Suffix

Morphological Parsing

The process of determining the morphemes (and their purposes) from which a given word is constructed.

Can be visualized in a tree diagram (morphology tree)





Morphological Rules

Language build more complex words out of morphemes via

- **Derivation**
- Compound
- Inflection

Derivation

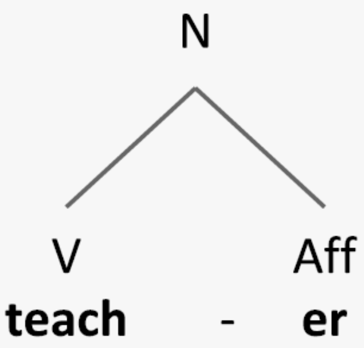
The process of forming a new word from an existing word by adding affixes

The meaning of the resulting word is different of its base

Very often there is a change in word category involved

Example

Teach	-er	teacher
Root	affix	resulting word
(verb)		(noun)



Morphological Rules

Language build more complex words out of morphemes via

- Derivation
- Compound
- Inflection

Compound

Combination of already existing words into new ones

There is no affixation but each of the parts can be assigned to a certain word category.

Example

N	+	N	➔	N	: lawn mover
P	+	N	➔	N	: up shot
N	+	V	➔	N	: blow dry
P	+	Adj	➔	Adj	: over grown
		Head			

Morphological Rules

Language build more complex words out of morphemes via

- Derivation
- Compound
- Inflection

Inflection

Modification of a word to express different grammatical categories such as tense, case, aspect, person, number, gender and mood

In English inflection is predominantly expressed by affixation

English has only eight inflection affixes

noun plural {-s}	He has three desserts.
noun possessive {-s}	This is Betty’s dessert.
verb present tense {-s}	Bill usually eats dessert.
verb past tense {-ed}	He baked the dessert yesterday.
verb past participle {-en}	He has always eaten dessert.
verb present participle {-ing}	He is eating dessert now.
adjective comparative {-er}	His dessert is larger than mine.
adjective superlative {-est}	Her dessert is the largest.

Inflection vs. Derivation

Derivation often changes the category of the base; inflection never does that.

Derivation changes the meaning of the base, inflection does that.

Derivation applies before inflection.

1. The farmer's cows escaped.
2. It was raining.
3. Those socks are inexpensive.
4. Jim needs the newer copy.
5. The strongest rower continued.
6. The pit-bull has bitten the cyclist.
7. She quickly closed the book.
8. The alphabetization went well.

Inflection vs. Derivation

Derivation often changes the category of the base; inflection never does that.

Derivation changes the meaning of the base, inflection does that.

Derivation applies before inflection.

1. The farmer's cows escaped.
2. It was raining.
3. Those socks are inexpensive.
4. Jim needs the newer copy.
5. The strongest rower continued.
6. The pit-bull has bitten the cyclist.
7. She quickly closed the book.
8. The alphabet-iz-ation went well.

Stemming vs. Lemmatization

Stemming

The process of reducing inflected or something derived words to their word stem

Example: cats ➔ cat

Morphological Parse of cats: cat + N + PL

Lemmatization

The process of grouping together the inflected forms of a word so that they can be analyzed as a single item, identified by the word's lemma or dictionary form.

Example: better ➔ good

01-03
Evaluation, Precision and Recall

01 Language and Morphology

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

Ground Truth

		predicted	
		true	false
actual	true	true positive	false negative
	false	false positive	true negative

Experiment

Let’s consider the following text corpus: BEETHOVENCORPUS

http://bit.ly/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.

Experiment

Task: Identify sentences that refer to Ludwig van Beethoven

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.

••• Experiment

Task: Identify sentences that refer to Ludwig van Beethoven

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.	

Ground Truth

Experiment

Task: Identify sentences that refer to Ludwig van Beethoven

Baseline Algorithm: Exact String Match with full name "Ludwig van Beethoven"

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.	

Identified **3** lines (1, 4, 8) as positive

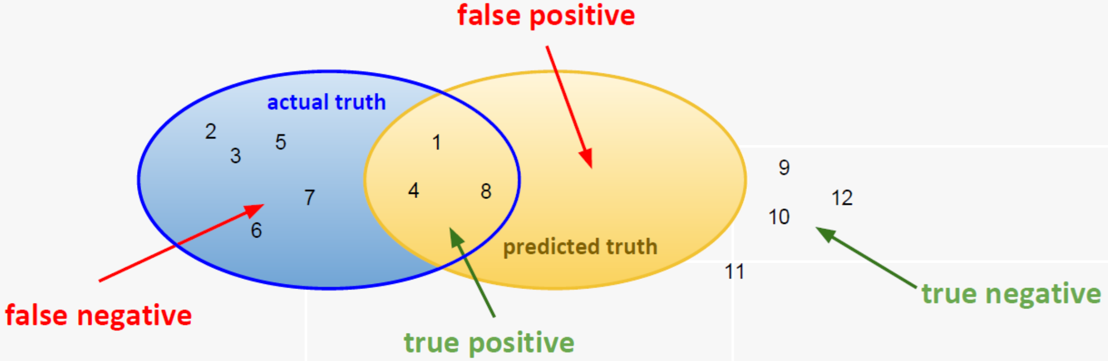
Identified **9** lines (2, 3, 5, 6, 7, 9, 10, 11, 12) as negative

Experiment

Baseline Algorithm: Exact String Match with 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

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

Recall

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

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

$$recall = \frac{3}{3+5} = 37.5\%$$

		predicted	
		true	false
actual	true	3	5
	false	0	4

Precision

Precision is the fraction of retrieved instances that are relevant

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

$$recall = \frac{3}{3+0} = 100\%$$

		predicted	
		true	false
actual	true	3	5
	false	0	4

F-Measure

F-Measure is a measure that combines precision and recall.

F_1-Measure is the harmonic mean of precision and recall.

$$F_1 = 2 \, . \frac{precision \, . \, recall}{precision + recall}$$

$$recall = 2 \, . \frac{100 * 37.5}{100+37.5} = 54.5\%$$

		predicted	
		true	false
actual	true	3	5
	false	0	4

Experiment

Task: Identify sentences that refer to Ludwig van Beethoven

Another Algorithm: Exact String Match with surname “Beethoven”

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.	

Identified **10** lines (1,2,3,4,5,7,8,9,10,11) as positive

Identified **2** lines (6, 12) as negative

Experiment

Another Algorithm: Exact String Match with surname “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 lines of it (12) are actual **negative** (true negative)
- 1 lines of it (6) are actual **positive** (false negative)

$$Precision = \frac{7}{10} = 70\%$$

$$Recall = \frac{7}{8} = 87.5\%$$

$$F_1 = 77.7\%$$

		predicted	
		true	false
actual	true	7	1
	false	3	1

01-04
Regular Expressions

01 Language and Morphology

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 Kleene (1956), used for text search first by Thompson (1968)

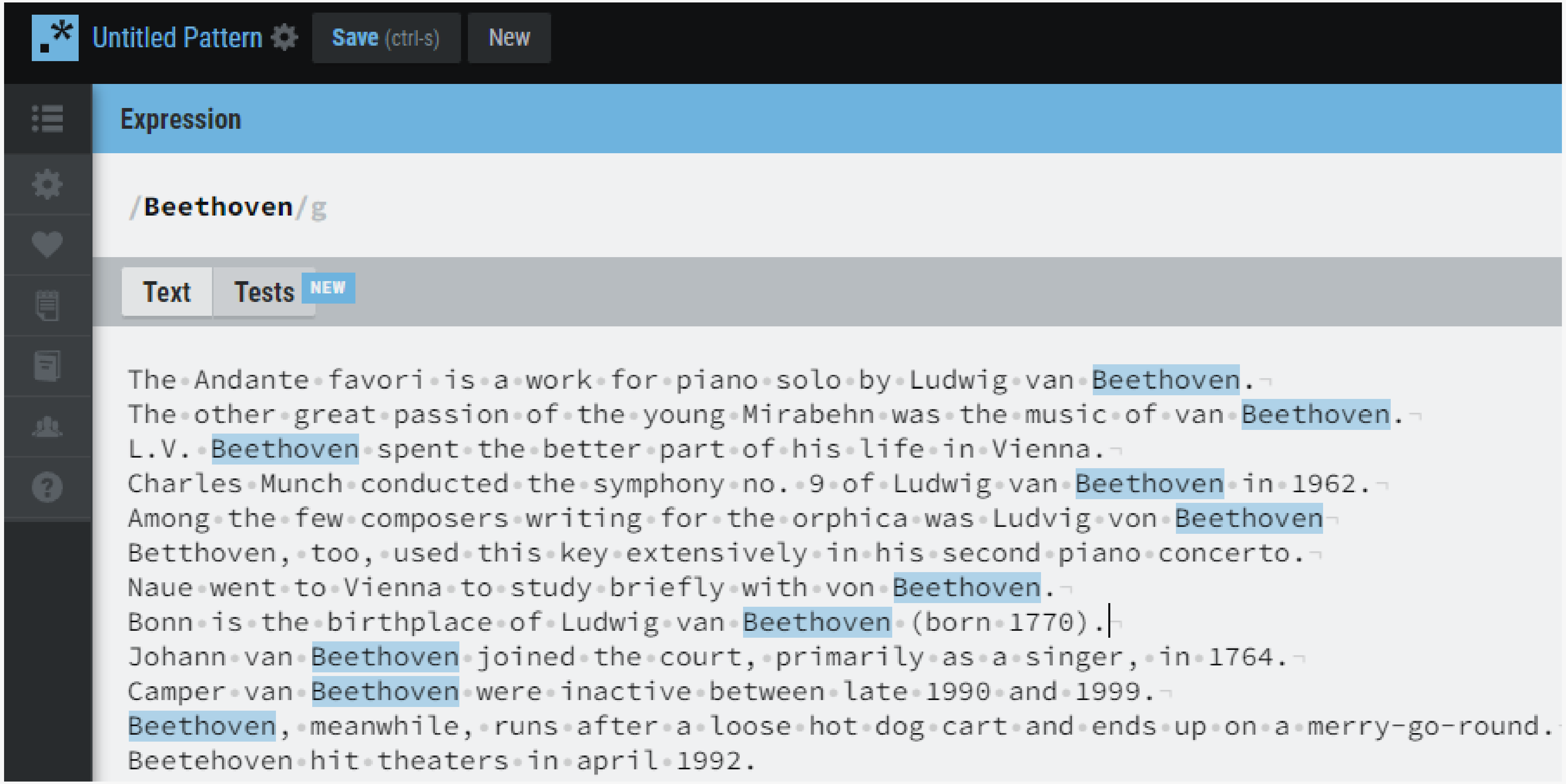
RE are an algebraic notation that specifies simple classes of strings

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

RegexR.com

/Beethoven/



```
/Beethoven/g
```

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

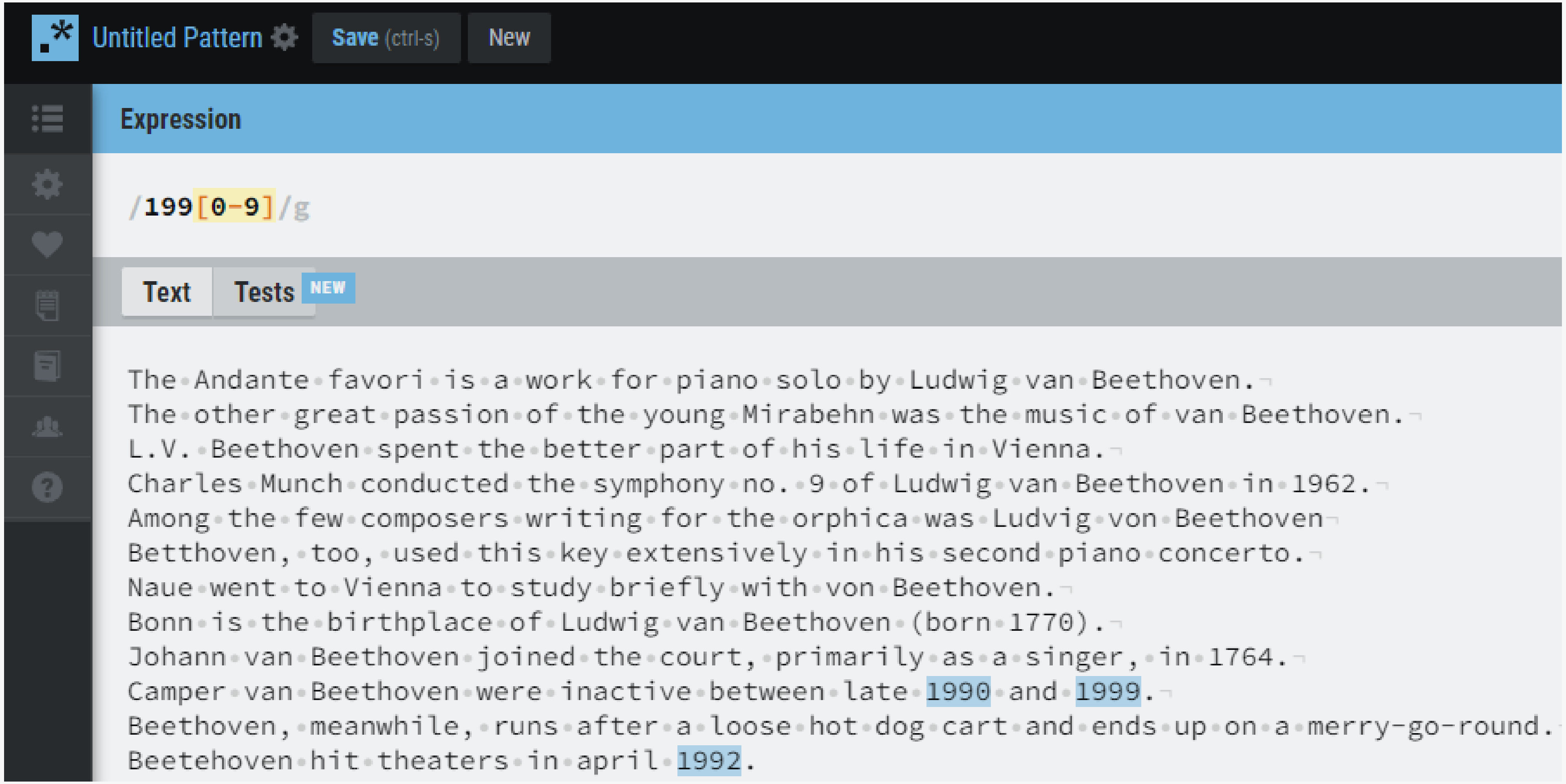
Disjunction

/199 [02]/

[0-9] any single digit

[a-z] any lower-case letter

[A-Z] any upper-case letter



```
/199[0-9]/g
```

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

Negation

/199 [ˆ2]/

[ˆ0-9] not a digit

[ˆa-z] not a lower-case letter

[ˆA-Z] not a upper-case letter

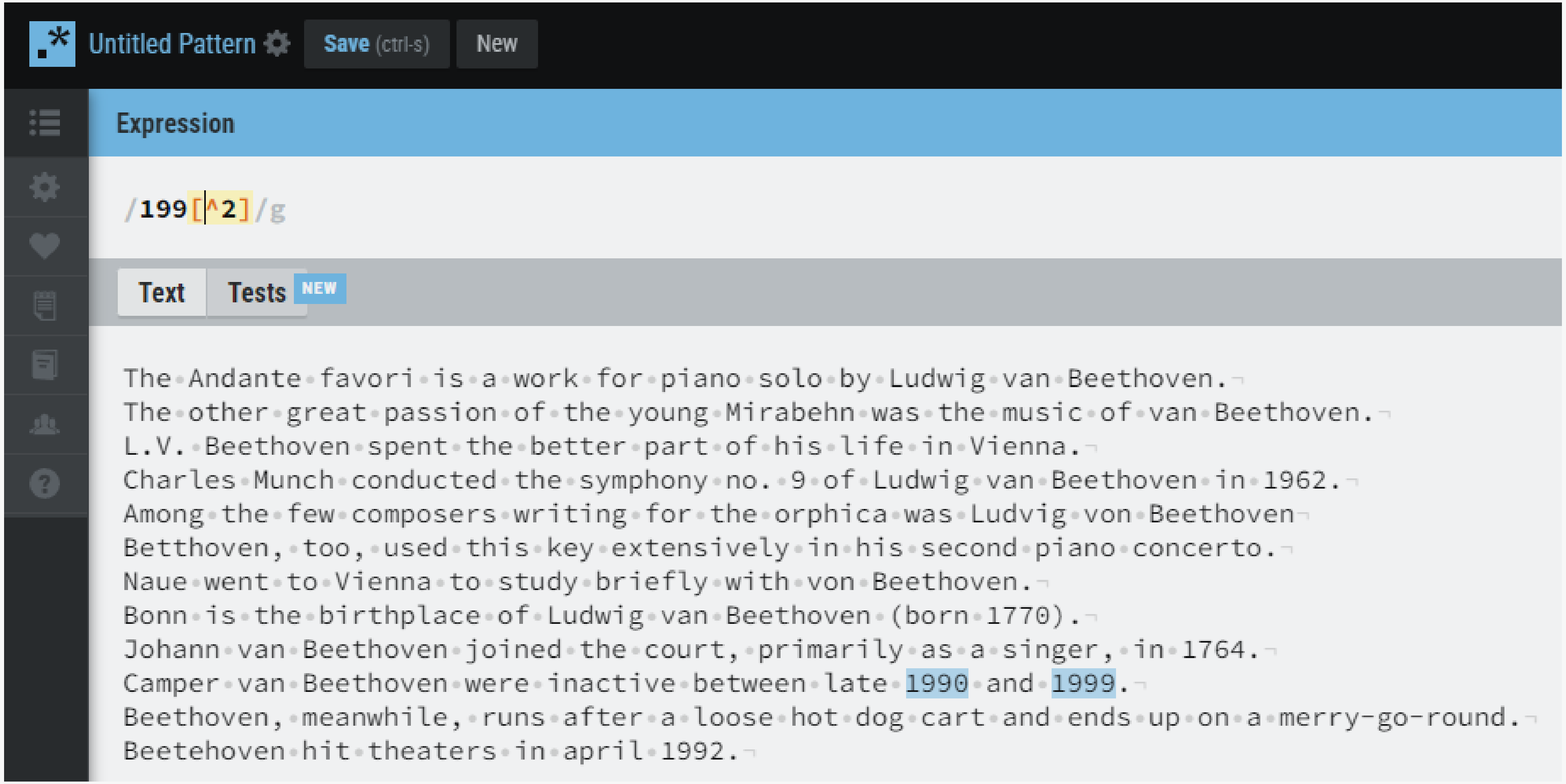
[ˆsS] neither s nor S

[ˆ\.] not a period

[eˆ] either e or ˆ

aˆb the pattern "aˆb"

Wildcard /199./



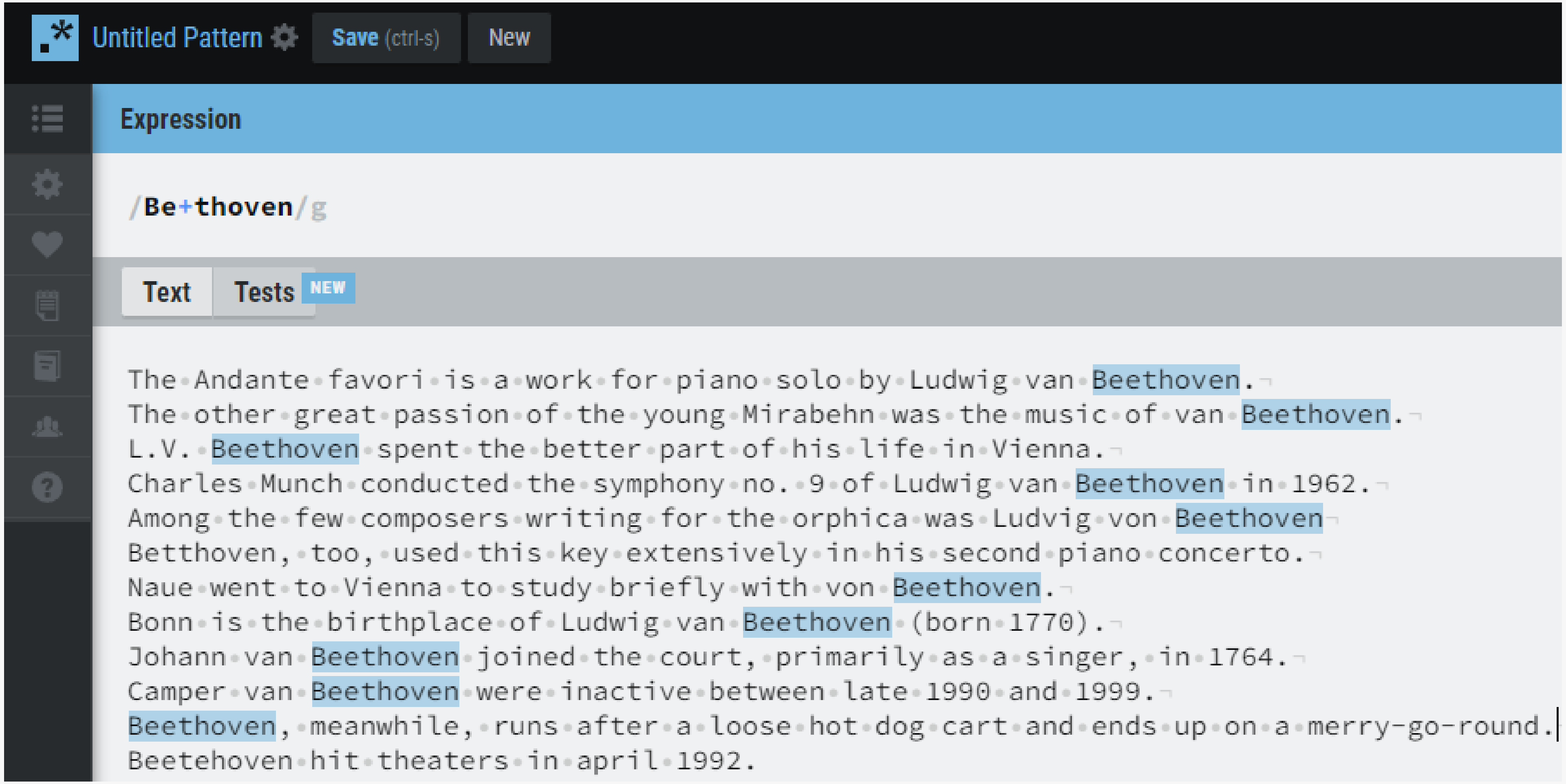
```
/199[^2]/g
```

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

Repetitive Pattern

/Be+thoven/

would also include
“Bethoven”,
“Beeethoven”,
“Beeeethoven”, etc.



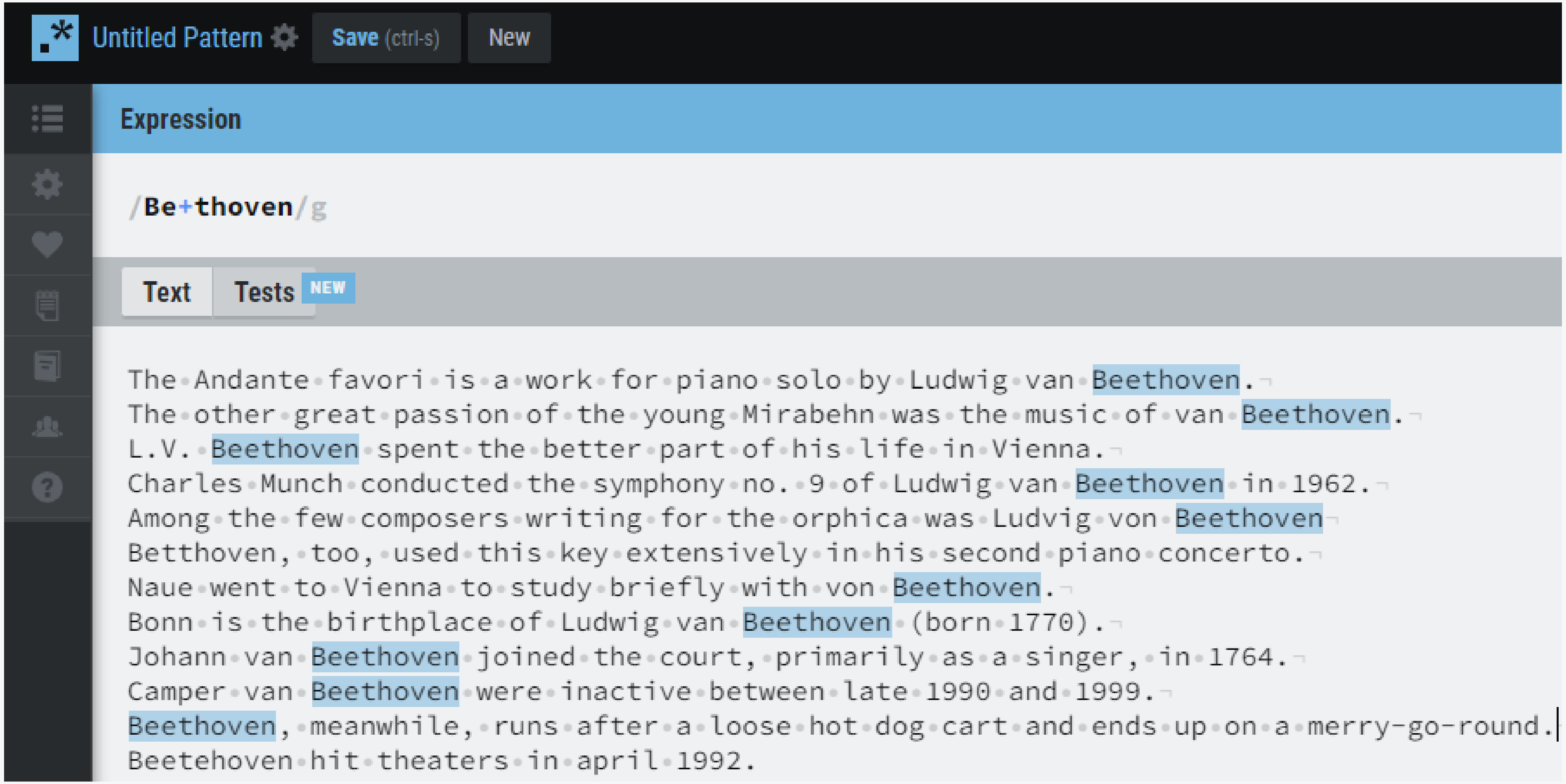
```
/Be+thoven/g
```

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

Optional and Repetitive Pattern

/Beete*hoven/

would also include
"Beethoven",
"Beeteehoven",
"Beeteeehoven",
"Beeteeeehoven", etc.



```
/Be+thoven/g
```

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

Regular Expressions

Anchor

ˆBeethoven	^ matches word only at start of the line
Beethoven$	$ matches word only at the end of the line

String Disjunction

Vienna|Bonn

Operator precedence

Parenthesis	( )
Counters	* + ? { }
Sequences and anchors	the ˆmy end$
Disjunction	|

Regular Expressions

Some characters need to be backslashed:

RE	Match	RE	Match
*	An Asterisk	\n	A Newline
\.	A Period	\t	A Tab
\?	A Question Mark	\,	A Comma

All functional characters that are to be used as ‘characters only’ in a pattern must be backslashed

Advanced Operators:

RE	Expansion	Match
\d	[0 – 9]	Any Digit
\D	[^0 – 9]	Any Non-Digit
\w	[a – z A – Z 0 – 9]	Any Alphanumeric + Underscore
\W	[^\w]	Any Non-Alphanumeric
\s	[\r \t \n \f]	Whitespace
\S	[^\s]	Non-Whitespace

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** occurrences 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

Synonyms and Variations

If we are searching for all occurrences of an entity in a text, we must 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. celluar phone
Related topics	e.g. cellphone video, cellular radio, phone carrier

Experiment

Task: Identify sentences that refer to Ludwig van Beethoven

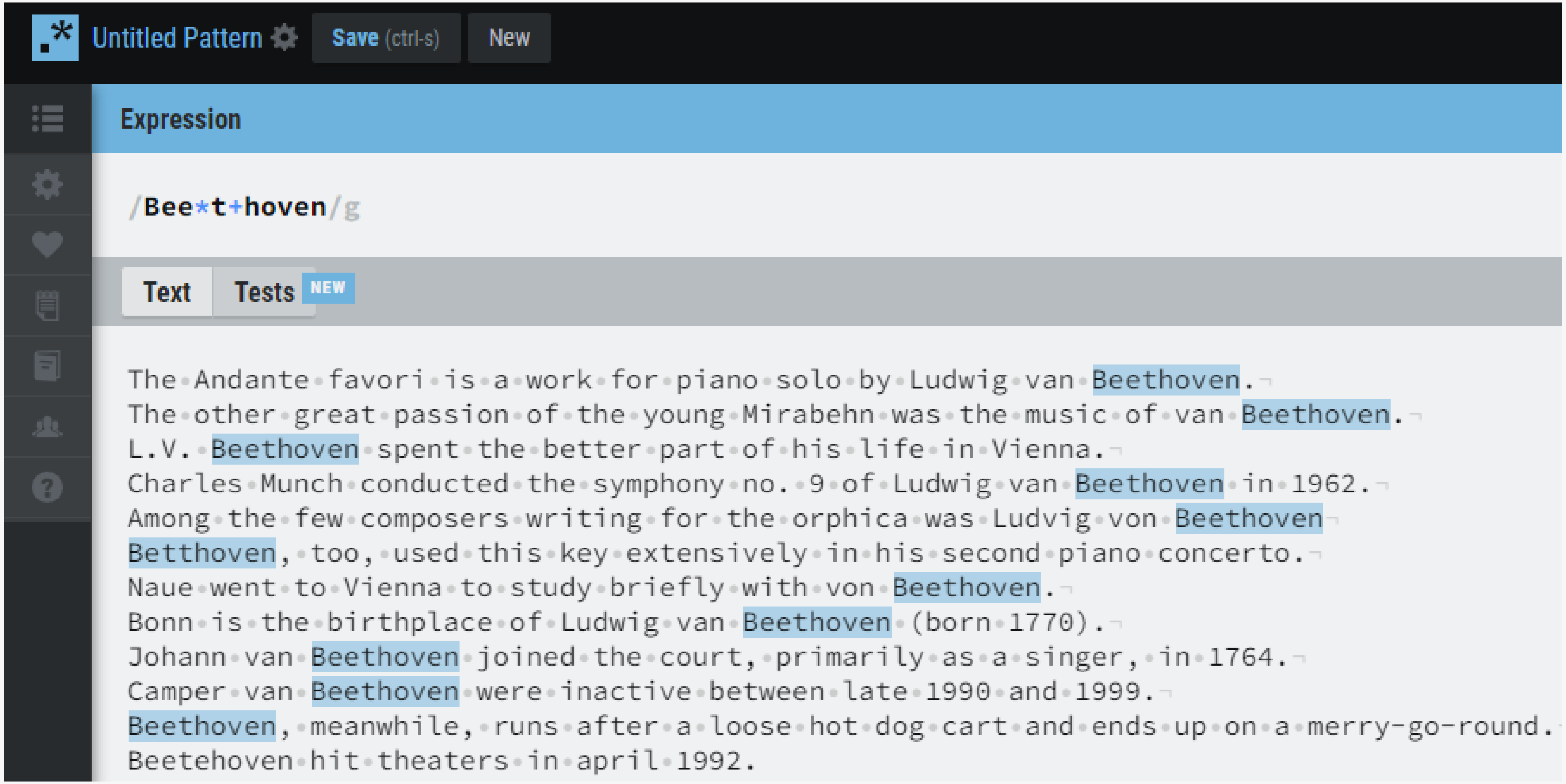
Another Algorithm: RE Match with surname “Bee*t+hoven”

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.	

Experiment

Task: Identify sentences that refer to Ludwig van Beethoven

Another Algorithm: RE Match with surname “Bee*t+hoven”



```
/Bee*t+hoven/g
```

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

Experiment

Another Algorithm: RE Match with surname “Bee*t+hoven”

Identified **11** lines (1,2,3,4,5,6,7,8,9,10,11) as positive

- 8 lines of it (1,2,3,4,5,6,7,8) are actual **positive** (true positive)
- 3 lines of it (9,10,11) are actual **negative** (false positive)

Identified **1** lines (12) as negative

- 1 lines of it (12) are actual **negative** (true negative)
- 0 lines of it are actual **positive** (false negative)

$$Precision = \frac{8}{11} = 72.7\%$$

$$Recall = \frac{8}{8} = 100\%$$

$$F_1 = 84.2\%$$

		predicted	
		true	false
actual	true	7	0
	false	3	1

Assignment

Can you obtain $F_1 = 100\%$?

If so,

will this be the “perfect” Beethoven Identifier?