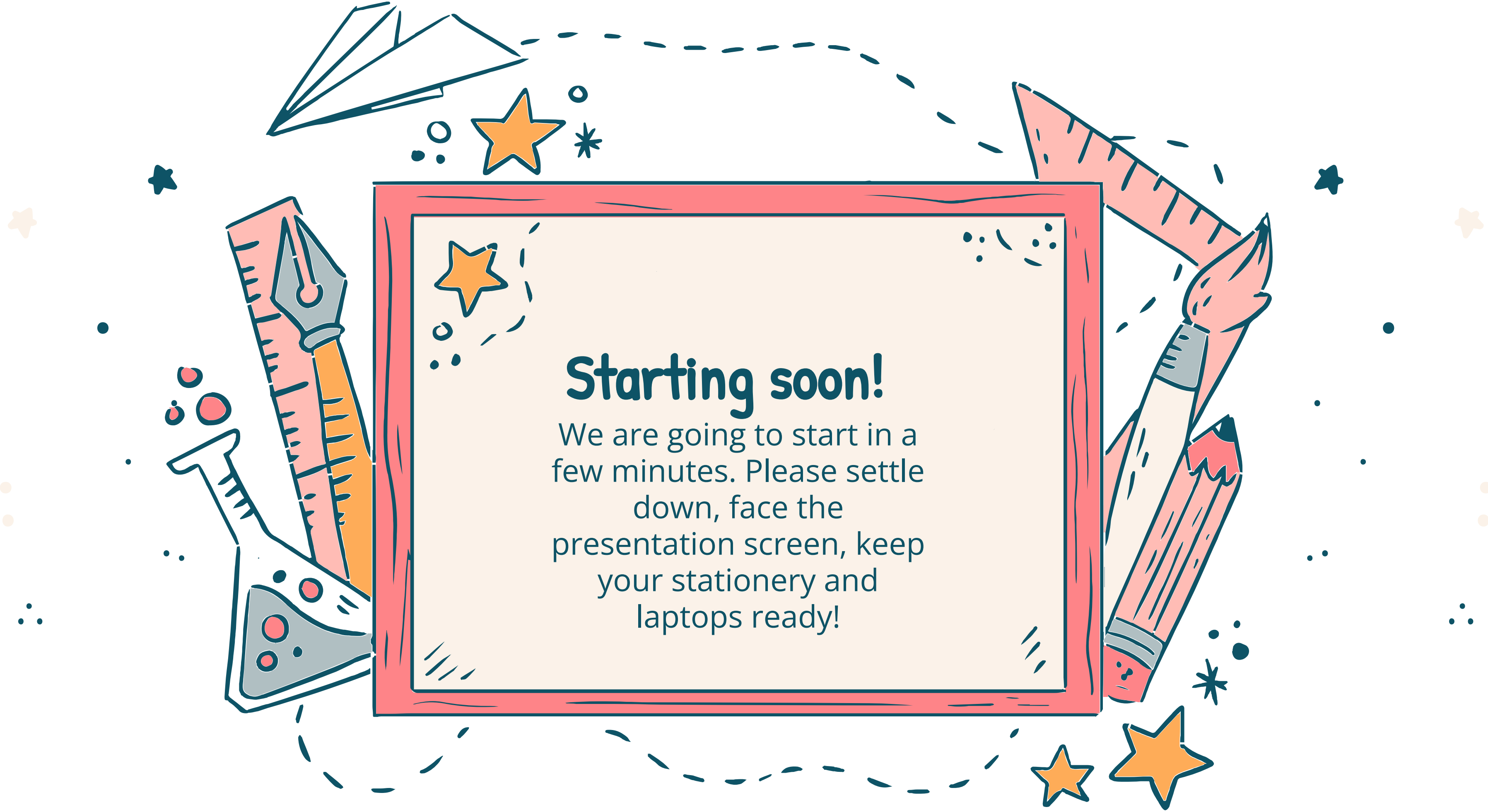
Introduction to Text Analytics

Exercise sessions!



Starting soon!

We are going to start in a few minutes. Please settle down, face the presentation screen, keep your stationery and laptops ready!

I am Pranav!

I have just started my job in Hamburg. I have been working in Hong Kong for a while and now moved here!

I do research on ethics in AI and multilingual AI (how can AI can understand mixture of languages?)

My hobbies: Guitar, Bedrotting, Music, Coffee……..

Use the office hours.

I am generally available from 4.30 - 5.30 pm everyday.

Feel free to come to my office about courses, research etc.

People come to my office about:

- Playing guitar with me
- Drinking coffee with me
- Ranting with me
- Watching tiktoks and memes with me

How to do well in your courses

Here is some advice I would give to younger students if they wish to do well in their courses.

Having been tested for many years of my life (with pretty good results), here are some rules of thumb that I feel helped me.

All-nighters are not worth it.

Sleep does wonders. Optimal sleep time for me is around 7.5 hours, with an absolute minimum of around 4hrs.

It has happened to me several times that I was stuck on some problem for an hour in the night, but was able to solve it in 5 minutes in the morning.

Before your tests

Create schedule of study, even if you dont stick to it. For me this usually involves getting an idea of everything I need to know and explicitly writing it down in terms of bullet points.

Consider all points carefully and think about how long it will take you to get them down. If you don't do this, there is a tendency to spend too much time on beginning of material and then skim through the (most important) later material due to lack of time.

Before your tests

Always try to look at previous tests BEFORE starting to study.
Especially if the past tests were written by the same professor.

This will give you strong hints about how you should study. Every professor has a different evaluation style. Don't actually attempt to complete the questions in the beginning, but take careful note of the type of questions.

Before your tests

Reading and understanding IS NOT the same as replicating the content.

Even I often make this mistake still: You read a formula/derivation/proof in the book and it makes perfect sense. Now close the book and try to write it down. You will find that this process is completely different and it will amaze you that many times you won't actually be able to do this!

Somehow the two things use different parts of the memory. Make it a point to make sure that you can actually write down the most important bits, and that you can re-derive them at will.

Before your tests

Study in groups, but at the END

Study alone first because in the early stages of studying others can only serve as a distraction. But near the end get together with others: they will often point out important pitfalls, bring up good issues, and sometimes give you an opportunity to teach.

Before your tests

Go to the prof before final exam at least once for office hours.

Even if you have no questions (make something up!) Profs will sometimes be willing to say more about a test in 1on1 basis (things they would not disclose in front of the entire class). Don't expect it, but when this does happen, it helps a lot. Does this give you an unfair advantage over other students? Sometimes. It's a little shady

But in general it is a good idea to let the prof get to know you at least a little.

Before your tests

Study well in advance.

Did I mention this already? Maybe I should stress it again. The brain really needs time to absorb material. Things that looked hard become easier with time.

Before your tests

If things are going badly and you get too tired, in emergency situations, jug an energy drink.

They work. It's just chemistry.

Before your tests

For things like this subject: Exercise > Reading.

It is good to study to the point where you are reasonably ready to start the exercises, but then fill in the gaps through doing exercises, especially if you have many available to you. The exercises will also make you go back and read things you don't know.

Before your tests

Make yourself cheat sheet.

Even if you're not allowed to bring it to the exam. Writing things down helps.

What you want is to cram the entire course on 1 or more pages that you can in the end tile in front of you and say with high degree of confidence "This is exactly everything I must know"

Before your tests

Study in places where other people study as well, even if not the same thing.

This makes you feel bad when you are the one not studying. It works for me

Places with a lot of background noise are bad and have a research-supported negative impact on learning. Libraries and Reading rooms work best.

Before your tests

Optimal eating/drinking habit is: T-2 hours get coffee and food.

For me, Coffee or Food RIGHT before the test is ALWAYS bad
Coffee right before any potentially stressful situation is ALWAYS bad.
No coffee at all is bad.

I realize the coffee bit may be subjective to me, but its something to think about for yourself.

Before your tests

Study very intensely RIGHT before the test.

I see many people give up before the test and claim to "take a break". Short term memory is a wonderful thing, don't waste it!

Study as intensely as possible right before the test. If you really feel you must take a break, take it about an hour before the test, but make sure you study really hard 30-45 minutes before the test.

During your tests

Look over all questions very briefly before start.

A mere 1-3 second glance per question is good enough. Just absorb all key words, and get idea of the size of the entire test.

During your tests

On test, do easy questions first.

Do not allow yourself to get stuck on something too long. Come back to it later. I skip questions all the time... Sometimes I can complete as little as 30% of the test on my first pass. Some questions somehow become much easier once you're "warmed up", I can't explain it.

During your tests

Always try to be neat on the test.

Surprisingly few people actually realize this obvious fact: A human being will mark your test. A sad human being gives low marks. I suspected this as undergrad student and confirmed it strongly when I was TAing and actually marking.

During your tests

Always BOX IN/CIRCLE the answer

Especially when there is derivation around it. This allows the marker to give you a quick check mark for full marks and move on. Get in the mindset of a marker.

During your tests

NEVER. EVER. EVER. Leave test early.

You made a silly mistake (I guarantee it), find it and fix it. If you can't find it, try harder until time runs out. If you are VERY certain of no mistakes, work on making test more legible and easier to mark. Erase garbage, box in answers, add steps to proofs, etc.

I have no other way of putting this-- people who leave tests early are stupid. This is a clear example of a situation where potential benefits completely outweigh the cost.

During your tests

Communicate with the marker.

Show the marker that you know more than what you put down. Ok you can't do a particular step, but make it clear that you know how to proceed if you did.

Don't be afraid to leave notes when necessary. Believe it or not the markers often end up trying to find you more marks-- make it easy for them.

During your tests

Use a highlighter

Highlight or underline important steps or words.

In that way, even if you write nonsense with highlighted keywords, you might get full points!

During your tests

Consider number of points per question.

Many tests will tell you how many marks every question is worth. This can give you very strong hints when you are doing something wrong.

It also gives you strong hints at what questions you should be working on. It is, of course, silly to spend too much time on questions worth little marks that are still relatively hard for you.

During your tests

If there are <5 minutes left and you are still stuck on some question, STOP.

Your time is better spent re-reading all questions and making absolutely sure you did not miss any secondary questions, and that you answered everything. You wouldn't believe how many silly marks people lose this way.

Test advice

The most important advice for now.

No one cares

The crucial fact to realize is that noone will care about your grades, unless they are bad.

For example, I always used to say that the smartest student will get 85% in all of his courses.

This way, you end up with somewhere around 4.0 score, but you did not over-study, and you did not under-study.

Advice

Your time is a precious, limited resource. Get to a point where you don't screw up on a test and then switch your attention to much more important endeavors. What are they?

Advice

Getting actual, real-world experience, working on real code base, projects or problems outside of silly course exercises is extremely important.

Are you thinking of applying to jobs? Get a summer internship. Are you thinking of pursuing graduate school? Get research experience!

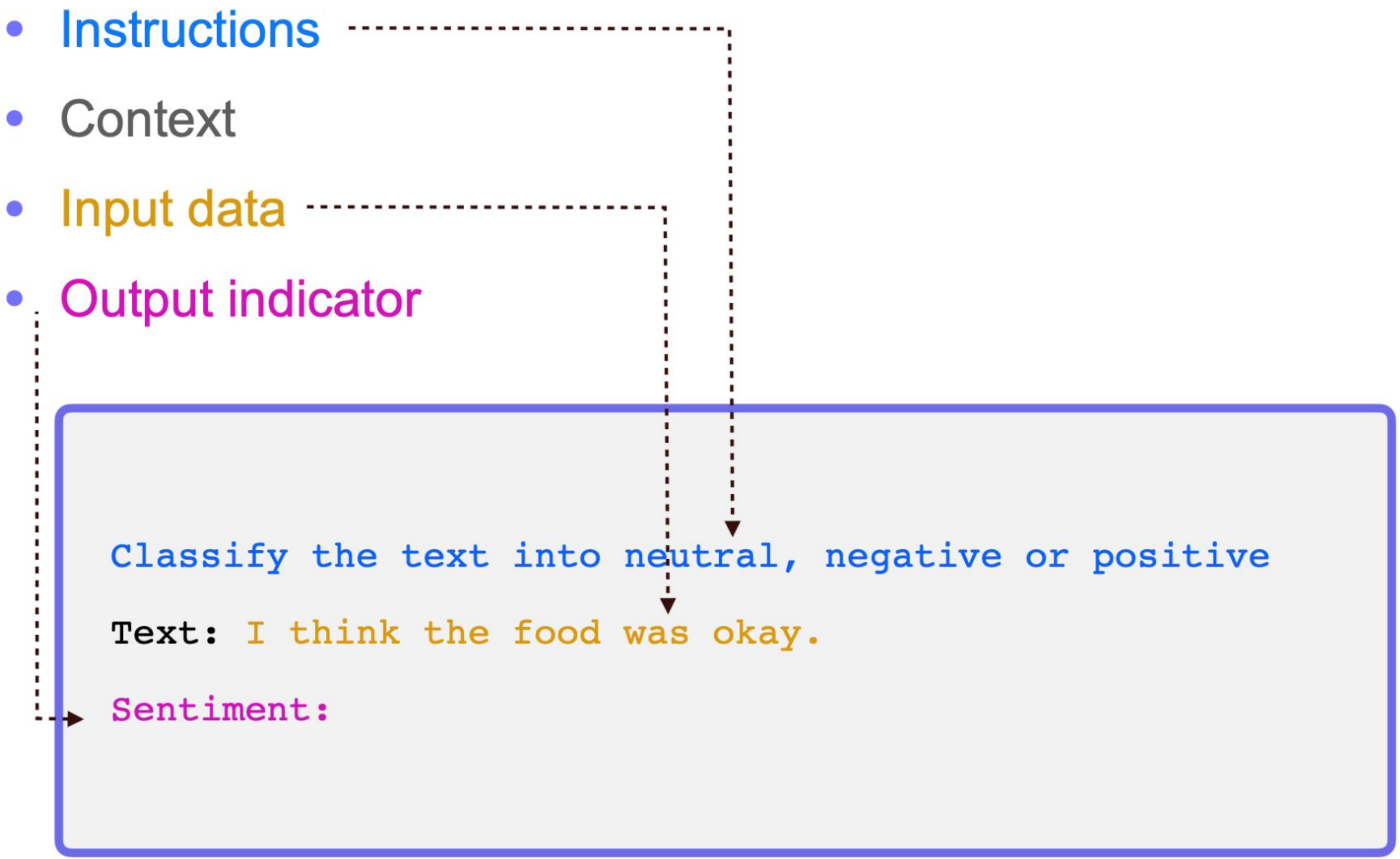
Advice

Get out there and create (or help create) something cool. Document it well. Blog about it.

These are the things people will care about a few years down the road.

Your grades? They are an annoyance you have to deal with along the way. Use your time well and good luck.

- A prompt is composed with the following components:



  - Instructions
  - Context
  - Input data
  - Output indicator

```
Classify the text into neutral, negative or positive

Text: I think the food was okay.

Sentiment:
```

Zero shot prompting

Prompt:

Classify the text into neutral, negative or positive.
Text: I think the vacation is okay.
Sentiment:

Output:

Neutral

Few shot prompting

Prompt:

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

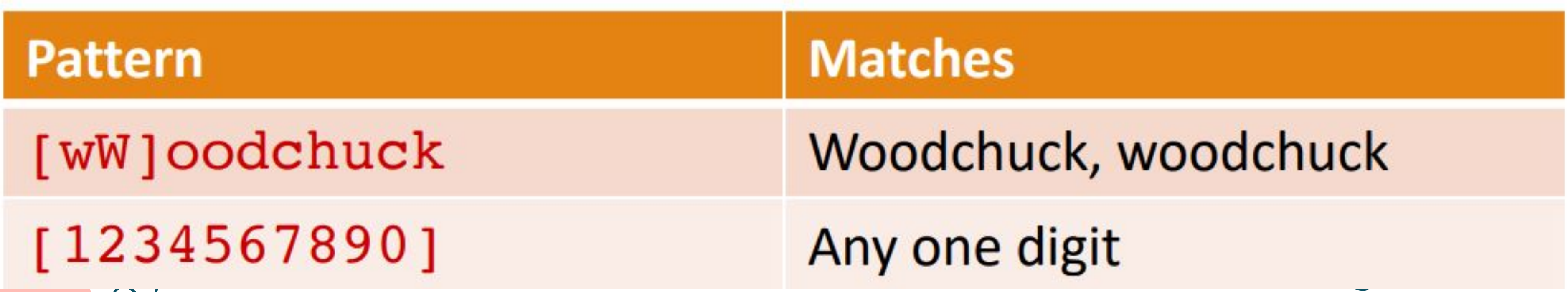
Output:

When we won the game, we all started to farduddle in celebration.

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

Regular Expressions

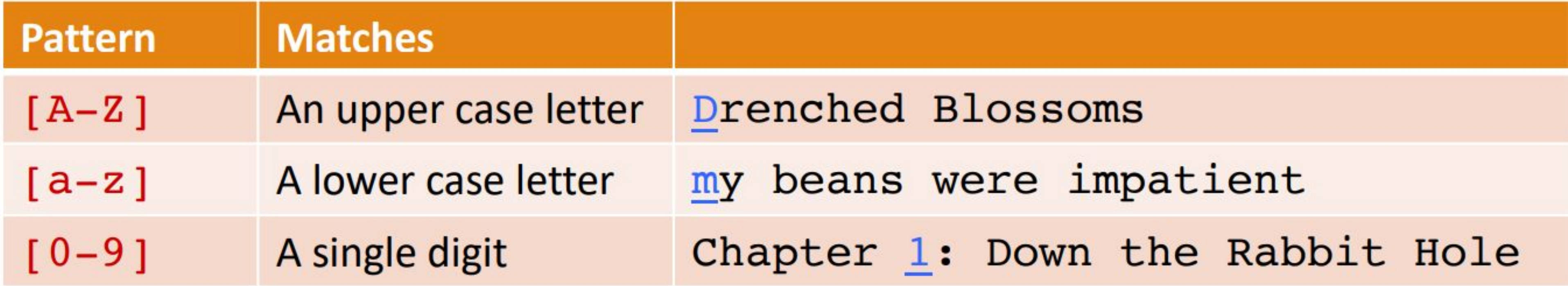
Letters inside square brackets []



Pattern	Matches
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any one digit

Regular Expressions

Ranges using the dash

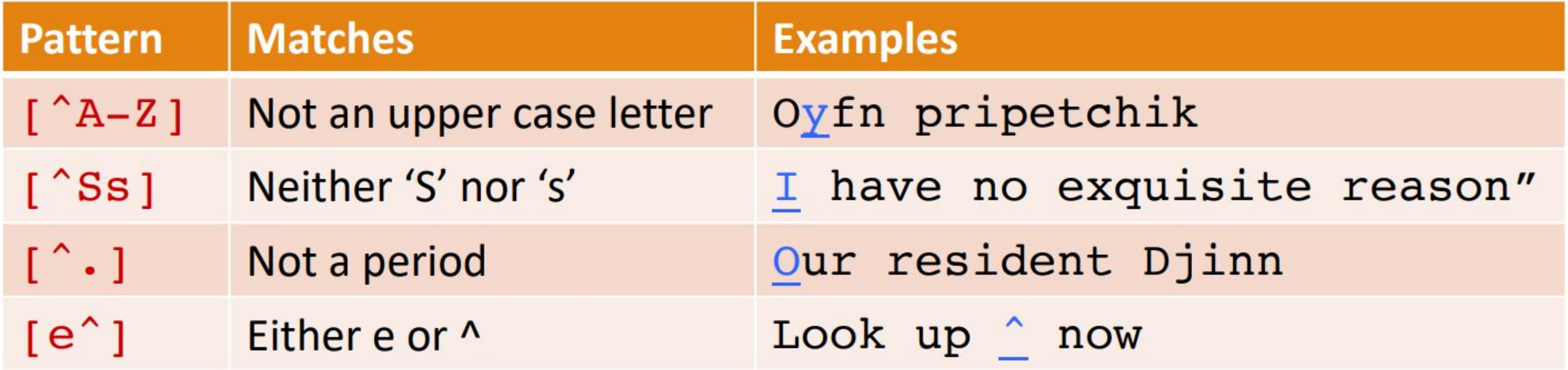


Pattern	Matches	
[A-Z]	An upper case letter	Drenched Blossoms
[a-z]	A lower case letter	my beans were impatient
[0-9]	A single digit	Chapter 1: Down the Rabbit Hole

Regular Expressions

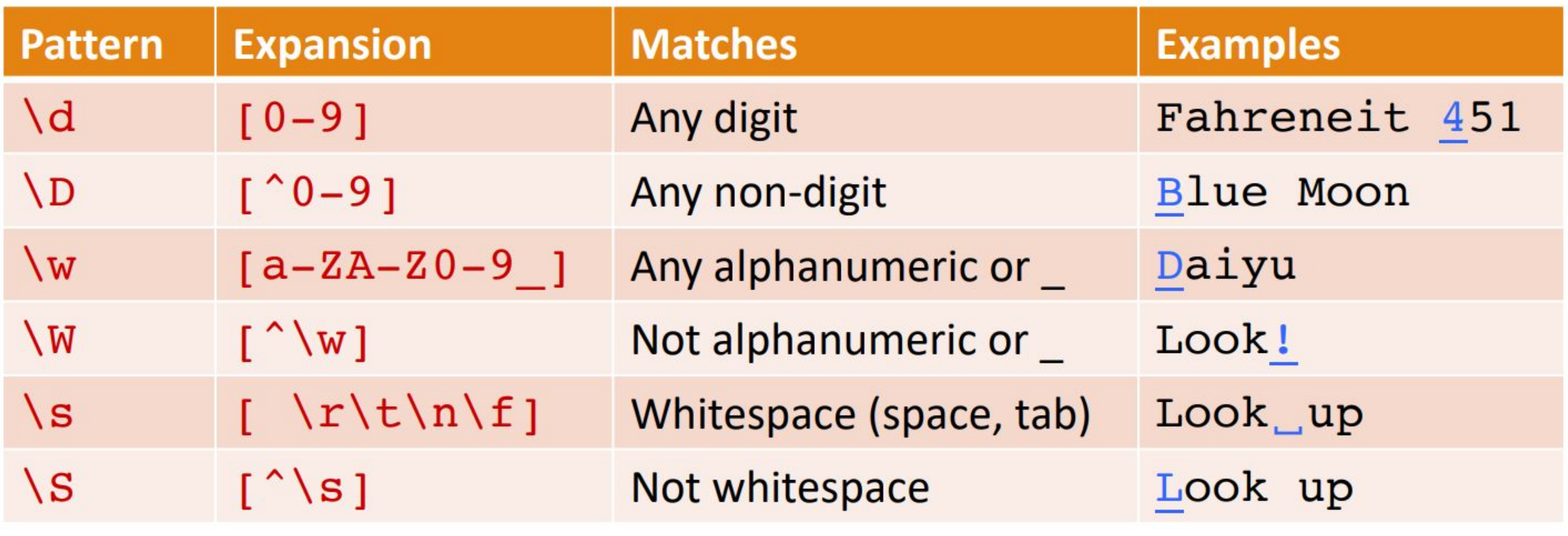
Carat as first character in [] negates the list

◦ Note: Carat means negation only when it's first in []
◦ Special characters (., *, +, ?) lose their special meaning inside []



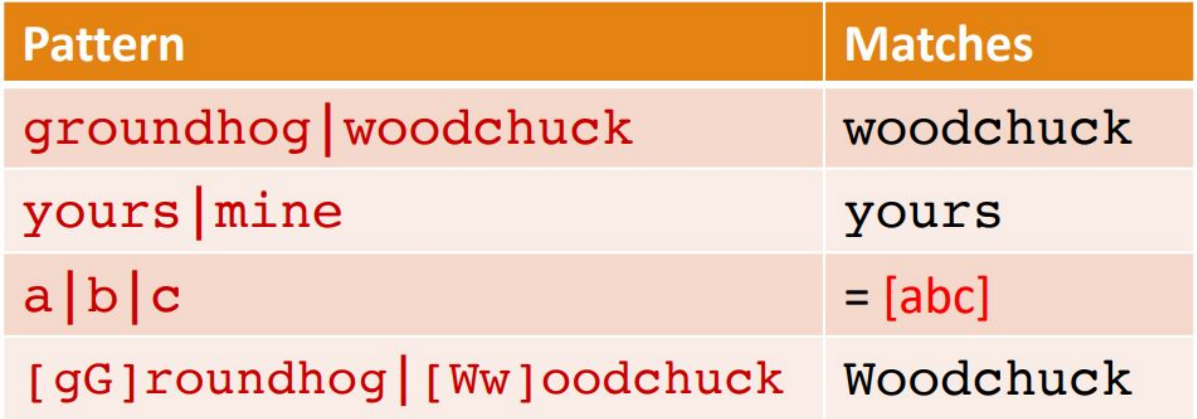
Pattern	Matches	Examples
`[^A-Z]`	Not an upper case letter	`Oyfn pripetchik`
`[^Ss]`	Neither ‘S’ nor ‘s’	`I have no exquisite reason”`
`[^.]`	Not a period	`Our resident Djinn`
`[e^]`	Either e or ^	`Look up ^ now`

Aliases



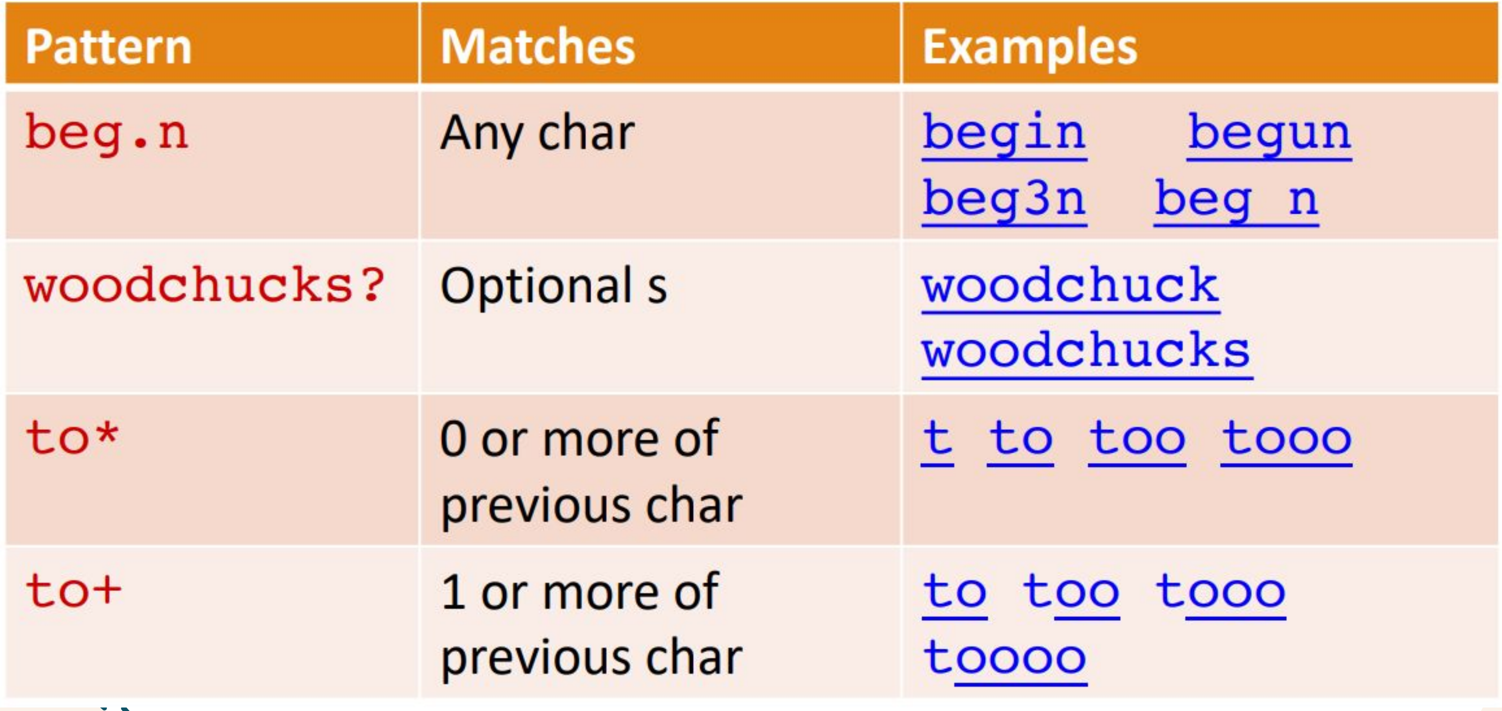
Pattern	Expansion	Matches	Examples
\d	[0-9]	Any digit	Fahreneit 451
\D	[^0-9]	Any non-digit	Blue Moon
\w	[a-ZA-Z0-9_]	Any alphanumeric or _	Daiyu
\W	[^\w]	Not alphanumeric or _	Look!
\s	[\r\t\n\f]	Whitespace (space, tab)	Look up
\S	[^\s]	Not whitespace	Look up

OR



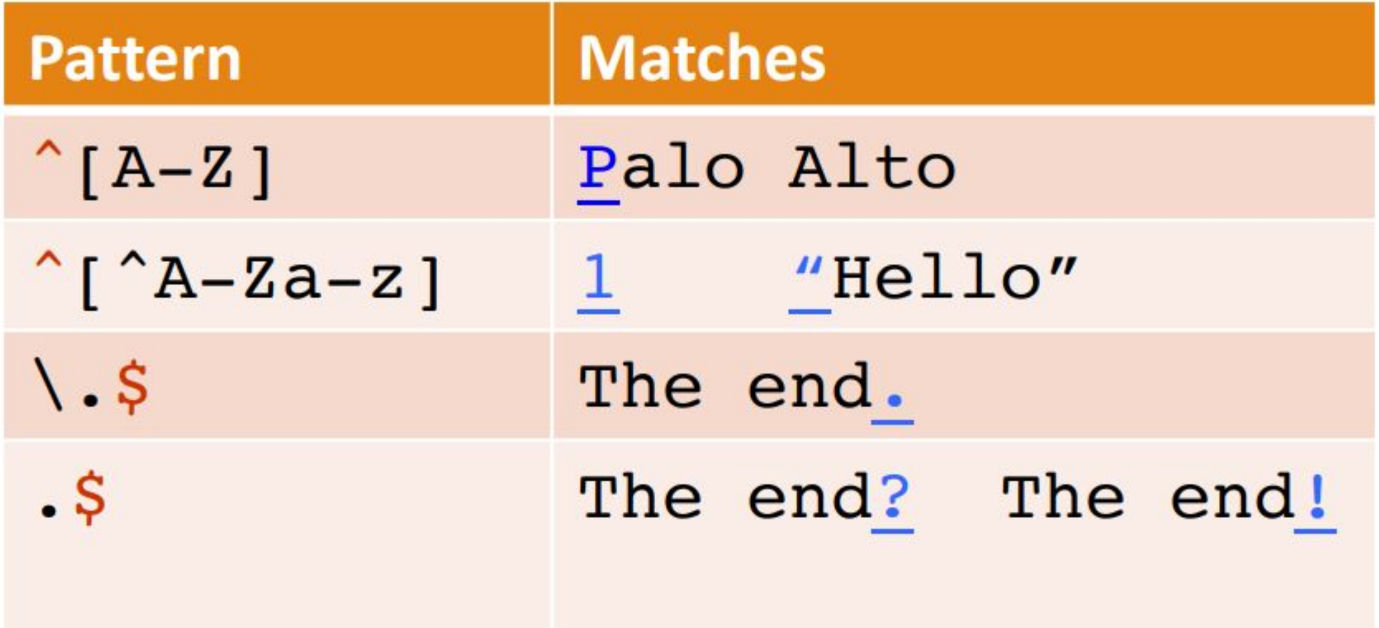
Pattern	Matches
groundhog|woodchuck	woodchuck
yours|mine	yours
a|b|c	= [abc]
[gG]roundhog|[Ww]oodchuck	Woodchuck

Wildcards



Pattern	Matches	Examples
beg.n	Any char	begin begun beg3n beg n
woodchucks?	Optional s	woodchuck woodchucks
to*	0 or more of previous char	t to too tooo
to+	1 or more of previous char	to too tooo toooo

Anchors



Pattern	Matches
`^[A-Z]`	`Palo Alto`
`^[^A-Za-z]`	`1` `“Hello”`
`\.$`	`The end.`
`.$`	`The end?` `The end!`

German Postal Code

- Pattern: ^\d{5}$
- Description: German postal codes consist of 5 digits. This pattern matches a sequence of exactly five numerical digits.

VAT Number (Umsatzsteuer-Id)

- Pattern: ^DE\d{9}$
- Description: German VAT numbers start with "DE", followed by 9 digits.

German Phone Number

Pattern: ^\+49\d{9}$

Description: Matches German phone numbers. Begins with +49 and is followed by 9 digits.

Email

```
^[^@]+@[^@]+\.[^@]+$
```

1. Start at beginning of string or line
2. Include all characters except @ until the @ sign
3. Include the @ sign
4. Include all characters except @ after the @ sign until the full stop
5. Include all characters except @ after the full stop
6. Stop at the end of the string or line

Collocations

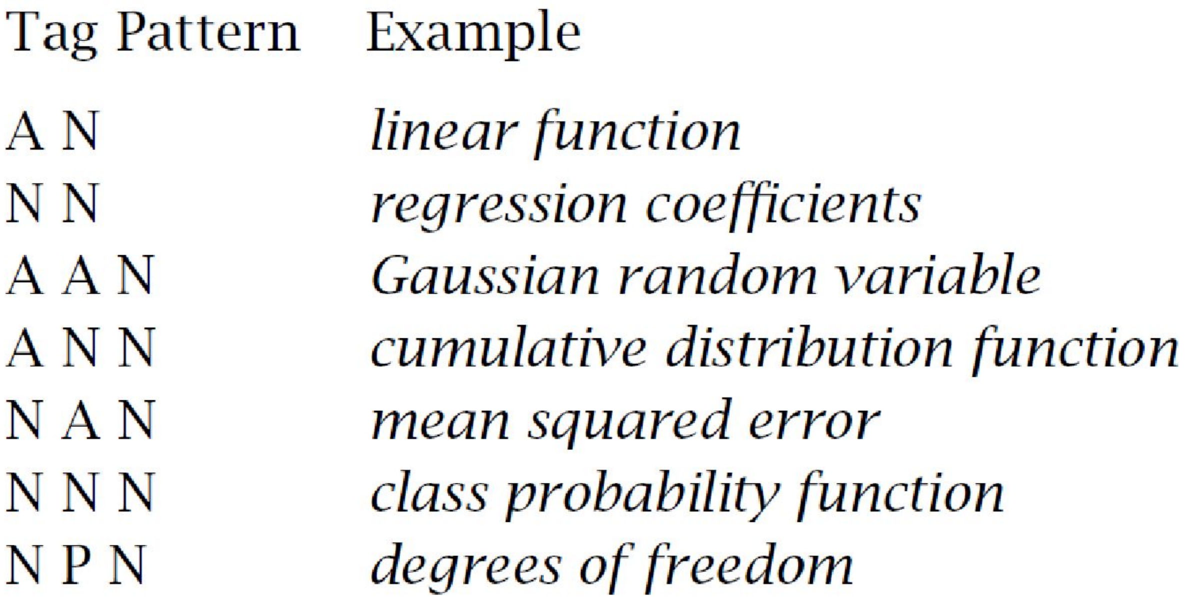
is an expression of two or more words that correspond to a conventional way of saying things.

– **broad daylight**

– Not **bright daylight** or **narrow darkness**

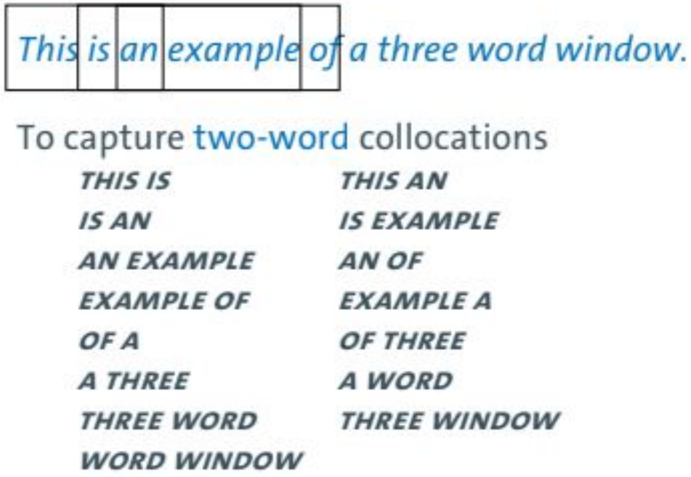
– **Big mistake** but not **large mistake**

Frequency Examples



Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Window example



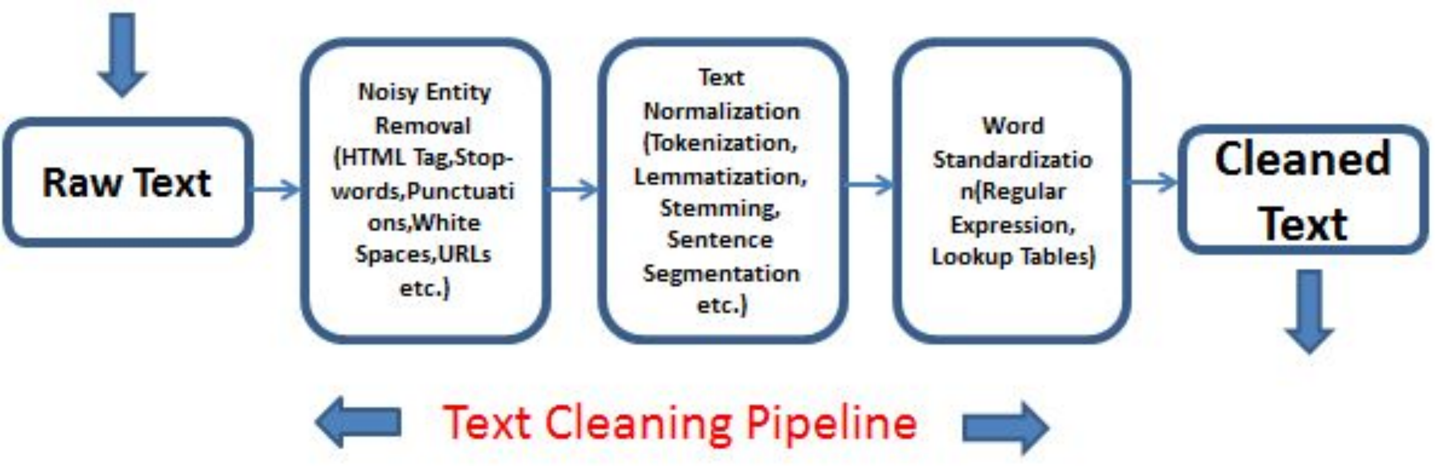
This is an example of a three word window.

To capture two-word collocations

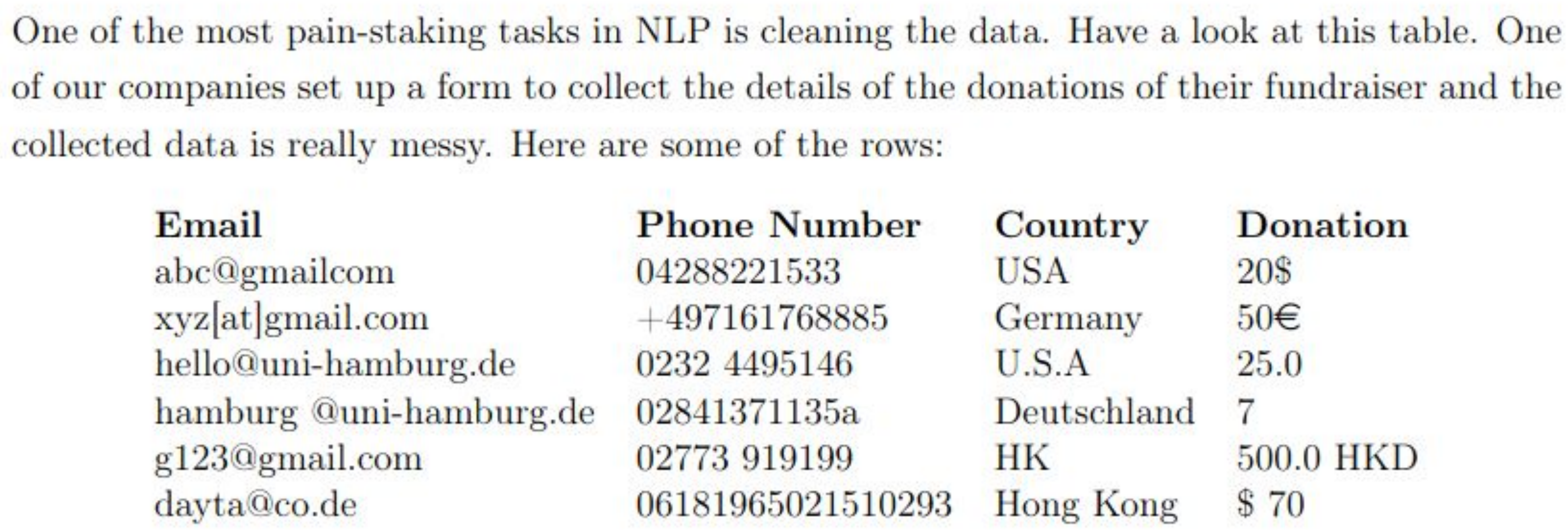
THIS IS
IS AN
AN EXAMPLE
EXAMPLE OF
OF A
A THREE
THREE WORD
WORD WINDOW

THIS AN
IS EXAMPLE
AN OF
EXAMPLE A
OF THREE
A WORD
THREE WINDOW

Text preprocessing


Raw Text
Noisy Entity Removal (HTML Tag,Stop-words,Punctuations,White Spaces,URLs etc.)
Text Normalization (Tokenization, Lemmatization, Stemming, Sentence Segmentation etc.)
Word Standardization(Regular Expression, Lookup Tables)
Cleaned Text
Text Cleaning Pipeline

Data Cleaning



One of the most pain-staking tasks in NLP is cleaning the data. Have a look at this table. One of our companies set up a form to collect the details of the donations of their fundraiser and the collected data is really messy. Here are some of the rows:

Email	**Phone Number**	**Country**	**Donation**
abc@gmailcom	04288221533	USA	20$
xyz[at]gmail.com	+497161768885	Germany	50€
hello@uni-hamburg.de	0232 4495146	U.S.A	25.0
hamburg @uni-hamburg.de	02841371135a	Deutschland	7
g123@gmail.com	02773 919199	HK	500.0 HKD
dayta@co.de	06181965021510293	Hong Kong	$ 70

Prompting

Write a one-shot prompt for a LLM (Large Language Model) to clean and normalize the country column.

In the following data, entries of countries entered by the customers are given. This data is not cleaned and standardized. Your task is to normalize this column.

Example: U.S.A and USA will be USA

Prompting

Write a one-shot prompt for a LLM (Large Language Model) to clean and normalize the donation column.

In the following data, entries of donations entered by the customers are given. This data is not cleaned and standardized. Your task is to normalize this column. Convert everything to EUR. Assume that everything customers have provided is in local currency.

Example: 20$ would be 18.50 EUR

Regex

The best way to avoid irregularities in the phone number column is to provide a regular expression-based validation. Write a regular expression for a German phone number.

```
^\+49[0-9]{9}$
```

Regex

Similarly, the best way to avoid irregularities in the email column is to provide a regular expression-based validation. Write a basic regular expression for an email.

```
^[^@]+@[^@]+\.[^@]+$
```

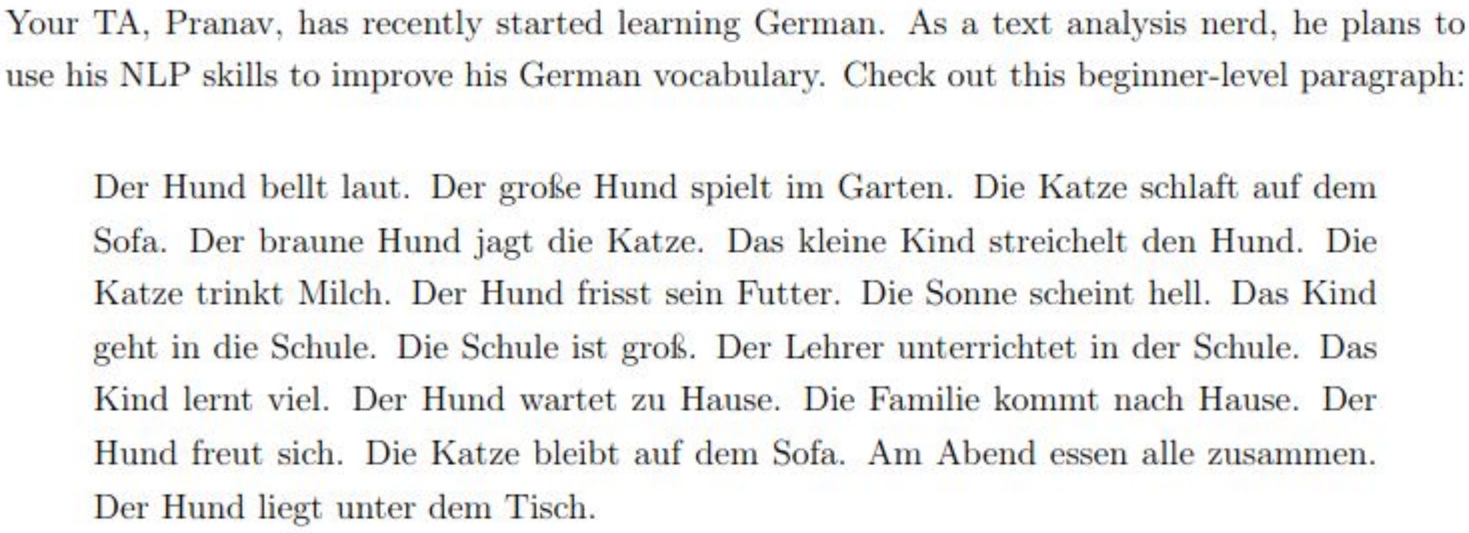
Formatting

Once the data is cleaned, you need to feed it to an API which accepts the data in JSON form. Write a prompt with a template to convert this dataset into JSON form.

Convert this data into a JSON form for the API input.
Example:

```
{
        email: {email_address},
        phone_number: {phone_number},
        ....
}
```

German Text analysis



Your TA, Pranav, has recently started learning German. As a text analysis nerd, he plans to use his NLP skills to improve his German vocabulary. Check out this beginner-level paragraph:

> Der Hund bellt laut. Der große Hund spielt im Garten. Die Katze schlaft auf dem Sofa. Der braune Hund jagt die Katze. Das kleine Kind streichelt den Hund. Die Katze trinkt Milch. Der Hund frisst sein Futter. Die Sonne scheint hell. Das Kind geht in die Schule. Die Schule ist groß. Der Lehrer unterrichtet in der Schule. Das Kind lernt viel. Der Hund wartet zu Hause. Die Familie kommt nach Hause. Der Hund freut sich. Die Katze bleibt auf dem Sofa. Am Abend essen alle zusammen. Der Hund liegt unter dem Tisch.

Regex

One of the biggest challenges for people who are learning German is to learn the grammatical gendered article of the words (like der, das, die). Write a regular expression to extract the German article followed by the word (for example: Der Hund, die Katze, das kind).

```
[D|d][er|as|ie] [a-zA-Z]+
```

Collocations

It is important for Pranav to learn the most frequent collocations to improve his vocabulary. Write a one-shot prompt and expected output for the LLM to find out at least 4 collocations here.

In corpus linguistics, a collocation is a series of words or terms that co-occur more often than would be expected by chance. Using the frequency window method, extract at least 4 good collocations in this paragraph.

An example would be: Der Hund

<para>

Collocations

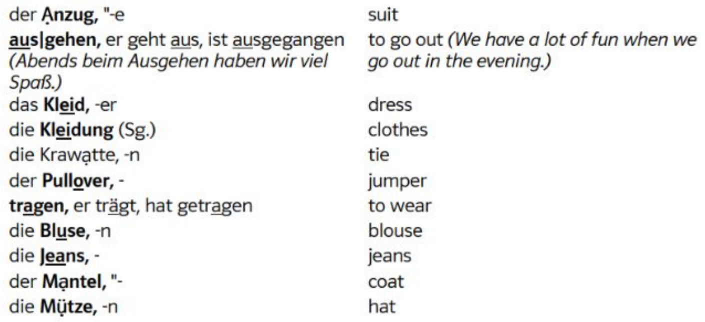
It is important for Pranav to learn the most frequent collocations to improve his vocabulary. Write a one-shot prompt and expected output for the LLM to find out at least 4 collocations here.

Output:

"Der Hund" appears 6 times
"Die Katze" appears 4 times
"Das Kind" appears 3 times
"Die Schule" appears 3 times

Prompting

Pranav will build a German-English vocab list. For that, he wants to extract nouns, verbs and adjectives from the German text and translate them into the English. He also wants to include variants of the verbs and plurals of the nouns in his vocab list. See the below picture of how it should look like. Write a few-shot prompt to output a similar list.



der **Ạnzug,** "-e	suit
aus|gehen, er geht aus, ist ausgegangen *(Abends beim Ausgehen haben wir viel Spaß.)*	to go out *(We have a lot of fun when we go out in the evening.)*
das **Kleid,** -er	dress
die **Kleidung** (Sg.)	clothes
die Krawạtte, -n	tie
der **Pullover,** -	jumper
tragen, er trägt, hat getragen	to wear
die **Bluse,** -n	blouse
die **Jeans,** -	jeans
der **Mạntel,** "-	coat
die **Mụ̈tze,** -n	hat

Prompting

From the given paragraph, make a German-English vocabulary. The first column will be German and the second column will be English. To do this, perform the following steps:

1. Extract the nouns (with articles), adjectives and verbs.
2. For nouns, include the plural forms.
3. For verbs, include the variants (present, past and perfect)
4. Arrange them in columns.

Example:

Das Kleid, -er	dress
Tragen, er tragt, hat getragen	to wear

Break

It’s time for a break now.

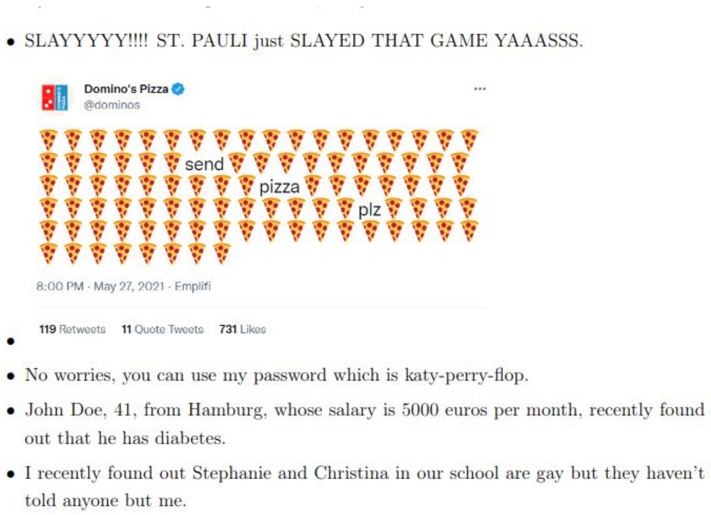
Have a break for like 10 minutes.

Then when you come back, sit in a groups of 3-4 folks and discuss around the question 3 (around 10-15 minutes).

Be respectful and give each other enough time.

Data cleaning

Check out these tweets and texts. What is the best way to clean this data? What tricks would you use to remove repeated letters, emojis and confidential information?



- SLAYYYYY!!!! ST. PAULI just SLAYED THAT GAME YAAASSS.
- Domino's Pizza
  @dominos

  🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕
  🍕🍕🍕🍕🍕🍕 send 🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕
  🍕🍕🍕🍕🍕🍕🍕🍕🍕 pizza 🍕🍕🍕🍕🍕🍕🍕
  🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕 plz 🍕🍕🍕🍕
  🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕🍕
  🍕🍕🍕🍕🍕🍕🍕🍕

  8:00 PM · May 27, 2021 · Emplifi

  119 Retweets 11 Quote Tweets 731 Likes
- No worries, you can use my password which is katy-perry-flop.
- John Doe, 41, from Hamburg, whose salary is 5000 euros per month, recently found out that he has diabetes.
- I recently found out Stephanie and Christina in our school are gay but they haven't told anyone but me.

Repeated Letters and Words

YASSSSSS -> YAS xxrep {5}

HIEEEEEE -> HIE xxrep {5}

please please please -> please xxwrep {3}

Confidential Information

Hide them with special tokens.

Use like:

[password]

[person]

Swear words

Many times, the texts might contain swear words. Would you consider removing those words? When would you do that, explain with examples.

If I am training the model from scratch, I won’t remove any swear words.

But for finetuning, say for like schools or professional settings, I will remove the swear words.

Blocklists

How are blocklists problematic? Would you still recommend going for blocklists in cleaning the data?

I will not use blocklists at all. Words have contexts! Unless I am sure that I have a specific use case for my AI, I won’t use blocklist.

Blocklist might cause problems in downstream applications like harming marginalized communities.