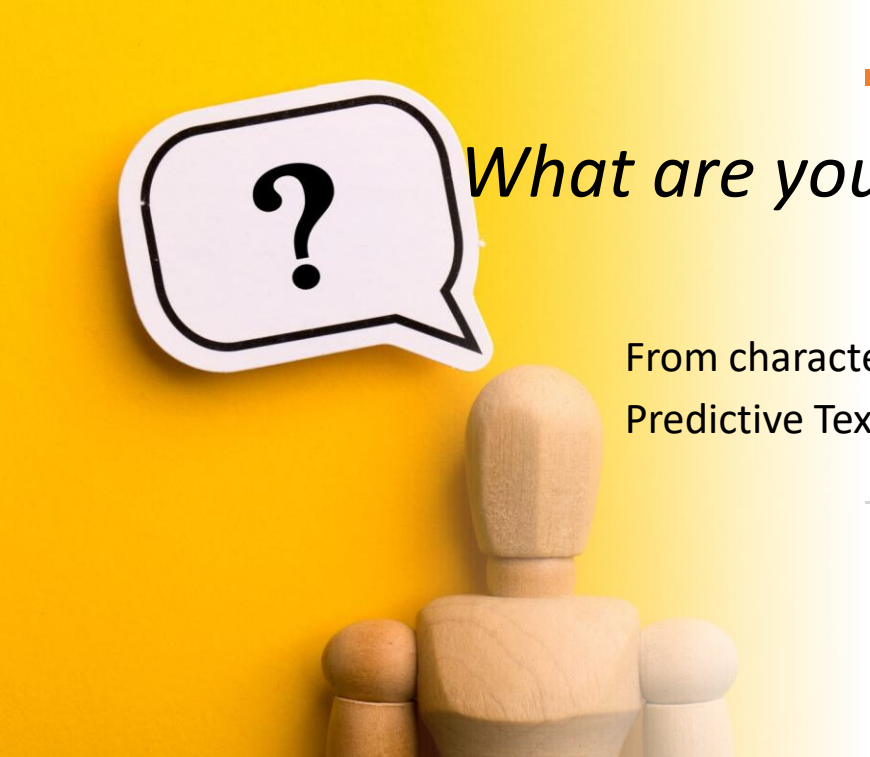


What are you going to say?

From characters to words

Predictive Text and n-grams

Confusion is a word we have invented for an order which is not understood.

. . .

It is the order of an accidental series of accidents accidentally conceived.

Tropic of Capricorn
Henry Miller

Information is Uncertainty and Surprise

- Shannon's discovery that "*information content*" can be interpreted in terms of a "*progressive reduction in the level of uncertainty*" has some far-reaching consequences that are still of importance today.
- Part of Shannon's research led to his investigating the notion of "*entropy and redundancy in spoken language*".
- Since "*character coding*" can lead to compression based on analysis of frequency what happens when we move from *characters* to ***words***?

We Expect Characters

- We saw that parts of "*character streams*" can have an element of "*predictability*".
- For example, 'Q' is (usually) followed by 'U';
- ' ' (SPACE) is (rarely) followed by 'X'.
- Although these are features of **written English**, similar patterns are found in other *Indo-European languages*.
- In fact, although the concept is very different, even in *Chinese*, *Japanese*, *Arabic*, *Sanskrit* etc etc.

Should We Expect Words?

Consider:

"*The cat sat on the . . .* "

Which of the following would "*most*" people use to complete the sentence? That is, to replace **. . .**

{*hat, rug, couch, sofa, mat, rat, ottoman, chaise-longue*}

Typically, the choice would be "*mat*".

Why?

Rhyme? But then why not "*hat*" or "*rat*"?

Sense? (you *sit* on *SOMETHING*). But then why not "*rug*", "*couch*", "*sofa*", "*ottoman*", "*chaise-longue*"?

FAMILIARITY: this is a standard child's reading exercise: "*ottoman*" and "*chaise-longue*" are obscure words.

Other Examples

a. The (rather irritating) habit of completing another person's sentence **BEFORE** they have finished speaking.
b. Possible *continuations* of text in editors such as Word.
c. Search term *suggestions* in Google.
d. Speech recognition & *automatic captioning*.

Important Differences

Notice that there are significant factors affecting (a) and the "*reclining attitudes of grimalkins*".

"*accurate*" prediction often depends on **sociological** and **cultural** background: *not statistical nuances*.

e.g. "*The cat sat on the mat*" will be recognised by most adults over a certain age, since it will be familiar from childhood.

Ending a speaker's sentence for them presumes a similar awareness of the speaker's topic and background.

Predictive Text

The cases in (b) (“smart” text editors) and (c) (search terms) are a little bit different.

These are less driven by “*shared cultural awareness*” (although this **does** feed in) and more by statistical observation.

[The feline adopts a sedentary posture on a “*mat*” because a **statistical analysis** suggests that “*most*” have “*been reported*” as “*mat sat*” rather than “*rat sat*” (or “*couch crouched*”).]

Speech Recognition – Automatic Captioning

- It is a *legal requirement* in the UK for (new) recorded and pre-recorded lectures provided to students to have *subtitles*/*closed captions*.
- Although most video embedding hosts (e.g., **YouTube**, **ms-streams**) offer a feature to extract *written* text from *spoken* content, the quality can be extremely variable and requires *manual editing* and *correcting*.
- This is a **VERY** tedious, tiresome, time-consuming, and labour-intensive process.

So why not just leave it?

- Some examples from recorded lectures:

a. "*Every exam has an **individual waiter***"

b. "*The next part will look at using **Foster Rivetts***"

c. "*We also consider advanced **May Tricks** operations*"

d. "*The lecture will be in the **Most Bad** Lecture Theatre*".

Notice that although the speaker's vocal cadences, inflections, and accent may influence transcription this is not always so.

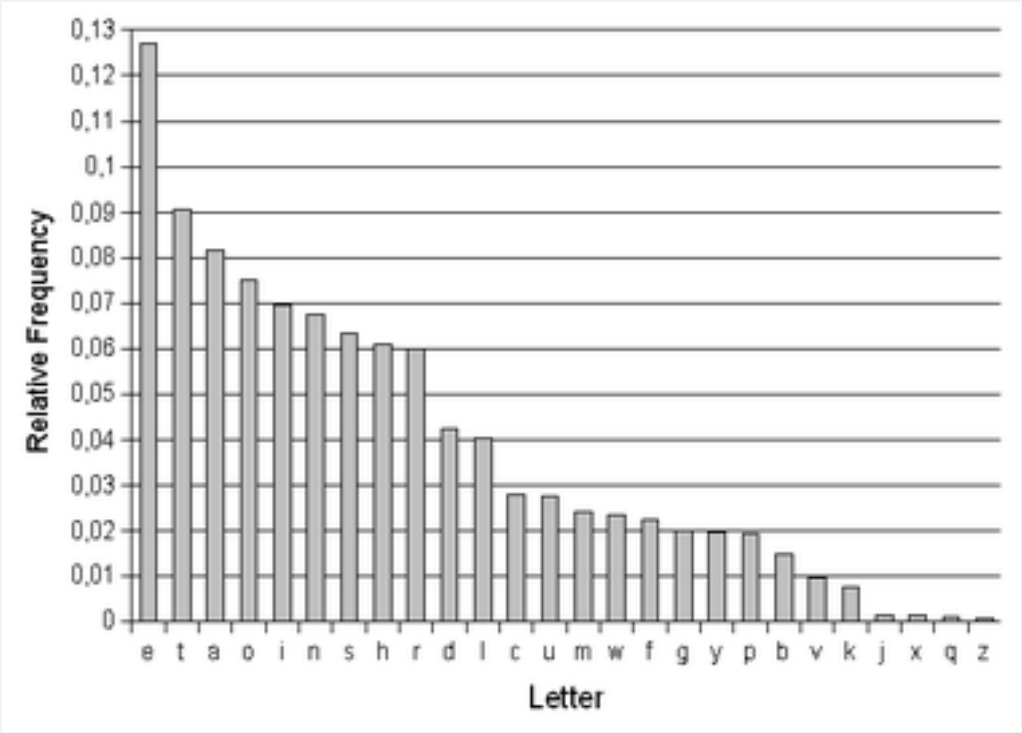
What was actually SAID

a. "*Every exam has an **invigilator***"
b. "*The next part will look at using **First Derivatives***"
c. "*We also consider advanced **Matrix** operations*"
d. "*The lecture will be in the **Musspratt** lecture theatre*".

A Difficulty

- It is, relatively, straightforward to construct *accurate* statistical data about *frequency* of **character** use.
- This can be done from comparatively "small" text samples, e.g., book chapters, newspaper articles, short stories.
- The *experimental claim* known as **Zipf's Law** (which is supported by several research studies) asserts:

"*The k'th most commonly used character in a language occurs roughly 1/k times as often in texts as the most frequently used*"


0,13
0,12
0,11
0,1
0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
Relative Frequency
e t a o i n s h r d l c u m w f g y p b v k j x q z
Letter

A Way with Words

- Although it is rather more demanding to analyze, a similar statistical model can be developed using **words** rather than **characters**.
- Further studies lead to the important concept of *n-gram language models*.
- We do not discuss these in depth here but simply introduce the basic elements.
- We also note that experimental studies support *Zipf's Law* when based on **words** as well as **characters**.

What's in a Word?

- In Indo-European languages (English, French, German, Italian, Spanish, Greek etc) a "*word*" may be interpreted as:

"*any sequence of characters from the language alphabet that is accepted by some authority*"

- For example, a *standard dictionary* (OED, Webster's, Larousse, Liddell-Scott, DRAE, DWB etc)
- These provide the basic units but to analyze usage frequency we need some **text corpus**: "*alphabets* are to *letter frequencies* as **words** are to their **use in texts**"

N-grams

- The idea behind **N-grams** is to use the *relative frequencies* of *sequences* of words as a guide to *prediction*, *interpretation*, *style analysis*, and, even, *creative writing*.
- Common choices of *N* are *N=2* (*bigram*) and *N=3* (*trigram*).
- An *N-gram* is a sequence $(w_1, w_2, w_3, \ldots, w_{N-1}, w_N)$ of *words*.
- Given some **text corpus** (with additional '*start*' and '*end*' sentence markers: $\{\langle s \rangle, \langle /s \rangle\}$)

N-grams and Relative Frequency

$$P\left[w_n \middle| w_{n-N+1} w_{n-N+2} \dots w_{n-1}\right] =$$

$$\frac{\#w_{n-N+1} w_{n-N+2} \dots w_{n-1}\, w_n}{\#w_{n-N+1} w_{n-N+2} \dots w_{n-1}}$$

- *Meaning*?

"*the probability of seeing the sequence* $w_{n-N+1} w_{n-N+2} \dots w_{n-1}\, w_n$ *is the number of times (in the text) that the sequence* $w_{n-N+1} w_{n-N+2} \dots w_{n-1}$ *is followed by* w_n *relative to the* ***total number*** *of times* $w_{n-N+1} w_{n-N+2} \dots w_{n-1}$ *is seen in the text*."

Small Example (N=2)

• 4 sentences

1. <s> In the beginning was the Word, and the Word was with God, and the Word was God. </s>
2. <s> The same was in the beginning with God. </s>
3. <s> All things were made by him: and without him was made nothing that was made. </s>
4. <s> In him was life, and the life was the light of men. </s>

John 1:1-4 (Douay-Rheims Edition)

[Note '*start*' (<s>) and '*end*' (</s>) sentence tokens. These are used to ensure adjustments are not needed for sentence lengths.]

Small Example (*N=2*) (*continued*)

- The following are all bigrams in this example:

  (the, word) ; (<s>, in) ; (God, </s>) ; (was, made) ; (the, beginning)

- #(the, word) = 3; #(the, beginning)=2; #(God, </s>)=2 ; #(<s>,all)=1
- #(the) = 7 ; #(God) = 3 ;
- These give relative frequencies

P[word | the] = 3/7 ; P[</s> | God] = 0.5;

P[beginning | the] = 2/7; P[all | <s>] = 0.25

- This, of course, is a very small example.

How is it used?

- Suppose N=2 (bigrams = “all sequences of pairs in a text”).
- First step: construct all bigrams and compute their *relative frequency* using the formula given.
- Prediction: “*when W is typed suggest X, where (W,X) has the* ***highest frequency*** *of bigrams starting W*”. (use a *threshold*)
- Creative writing: “*using an author’s* ***corpus*** *of written work, perform an analysis of bigram frequency. Use this to guide* ***random selections*** *of words to parody writing style*”

How is it used?

- Stylometry: "*compute an author's '**writing profile**' by forming a view of their use of specific bigram combinations*"
- In *plagiarism detection*, often N-grams building on *characters* rather than *words* are used. If a large enough item of text (for example project dissertation) is being analysed, then inconsistencies in profiles over "*text windows*" may indicate multiple authors and provide evidence of collusion and/or plagiarism.

Summary

- Natural Language Analysis is one significant application study within *Data Science*.
- Although the methodology offered by N-gram use has moved on and is now quite sophisticated, its initial development offers strong techniques.
- N-gram packages and *Natural Language* tools have been developed within **Python**. One of the most important being The Natural Language Toolkit
- **Stylometric Analysis** has been used to uncover *fraudulent practices* and has been considered as a tool to deal with *Generative AI abuses*.