Introduction to SketchEngine and WebBootCat as Corpus research

Professor Eric Atwell: So in this video, I'm going to be showing you the Sketch Engine web tool, which is used for text analytics and corpus linguistics research. Corpus linguistics is the specialist field of language or linguistics science based on using a corpus or text data set. So it is another word for text data analytics. Going to look particularly at the WebBootCaT tool within Sketch Engine, which allows you to collect a corpus of your own from the web.

So we're going to have a look at how to use this to collect and analyze, get some results out from the text data. Get a brief look at some research at Leeds University, the Leeds Internet Corpora, to collect very large corpuses or corpora. And then we're going to look at, again briefly, the Association for Computational Linguistics Special Interest Group on Web As Corpus, or ACL SIGWAC for short.

You should read Adam Kilgarriff's paper on the Sketch Engine 10 years on. Adam Kilgarriff was the researcher who founded Sketch Engine in the first place. And also, in the Minerva page, there's some links to various web pages to look at.

So what is Sketch Engine? It's an online tool, so you have to log in to use it. It's a commercial service, but we have free access for Leeds University researchers at the moment. Well, I say free. It's paid for, but you have it free. And it could do various things such as a word sketch, a summary of a word's behavior in collocations. Collocations is what words appear next to.

A concordance is an example of a particular word or lemma or phrase that appears repeatedly in the corpus, and it shows you some examples of its use in context. A distributional thesaurus shows comparing the concordance of this word with the concordance of other words or other words have got similar concordances. In other words, what other words appear in the same sorts of context as this word? And if they appear in the same similar context, they probably have similar meanings.

Can also have a parallel corpus. If you've got a bilingual corpus, a corpus and its translation into another language, then you can line up a word and its translation in various examples. Then we're going to look at WebBootCaT, which is a special tool for creating a specialised corpus of your own choice from the web. And once you've got that, then you can use terminology extraction to extract words and multi-word terms which are characteristic of that corpus and therefore, probably specialist terms. And there's more stuff too.

So let's have a look at some of the examples of these. We'll go into Sketch Engine later on and see some actual examples, or you can do this for yourself. But here's some examples from the paper by Adam Kilgarriff. This is a word sketch for the word catch. What sorts of words appear before and after it? So you can get to see what sorts of meanings catch has, its collocational behavior.

And this works not just for English, but other languages too. So I don't have to be able to read Arabic to be able to apply Sketch Engine to, given a particular word, find out what sorts of words come before it or what patterns it appears in. And the same for Chinese.

And so within Sketch Engine, they have a Gigaword corpus, a billion words of Chinese collected from the web using WebBootCaT. And the same thing for Arabic and same thing for other languages too. And we can see, for this Chinese word, typically where is it the object are subject modifiers and so on? What other grammatical functions it has.

Here's an example of a concordance. So a particular word or phrase, you can get from the corpus lots of examples of it so you can figure out what it means. If you're building a dictionary, and there's a phrase in English caught something or something or other pants, caught something pants, and we see lots of examples of caught with his pants down or caught with your digital pants down. Caught with your pants down as an idiom meaning you've been caught in a compromising situation with your pants down. Caught red-handed is another phrase meaning the same thing.

This is another use of distributional patterns. So within the corpus, the word tea, if you look at the collocations for tea, something like the pattern you saw before. And then look at other words which have similar distributional patterns. So coffee, the sort of patterns that coffee comes in are similar to the sort of patterns that tea comes in.

Therefore, tea and coffee probably mean similar things. And we've got other examples like drink, juice, chocolate, wine, and so on. And as you go down the list, they get less similar. So for example, vegetable apparently means the same as tea, and that's because you can have tea cup and vegetable cup and so on.

If you have a corpus in Chinese and the translation's in English-- and there are some corpora which are parallel corpora. For example, Opus 2 is one within Sketch Engine where we have, for a whole lot of documents, the English and the translated Chinese and the translation is in other languages too, and you could choose a word like smile and get the Chinese translations for smile, which lets you see what sort of words appear before and after smile in English and the equivalent in Chinese.

And the thing you're interested in for your coursework is WebBootCaT, which allows you to create a corpus from a web page. And in the paper, we have a slightly different look and feel to the

interface, remembering the paper is now a few years old. But it has the same functionality as we're about to see. So you create a corpus by choosing a name, choosing a language, choosing various other parameters.

Like, one way of doing it is by giving a list of URLs. But we don't know in advance what the URLs are. But you can give a list of seed words, and seed words are words or phrases that you think are going to be typical of your topic area and various other parameters.

And having got your corpus, we have terminology extraction. What this does is it compares the words in your corpus against the words in a standard English corpus, for example, the British National Corpus, and it finds words which are more frequent, relatively speaking, in your corpus than in the standard corpus. So if, for example, we had a volcanology corpus of volcan terms, then these are words which are more common in the volcanology corpus compared to, let's say, the British National Corpus.

Other things you can do in Sketch Engine too. You can extract word lists, that is words which are particularly frequent in your corpus. It's not the same thing as the terms we just got out because the word list is just a frequency list from your corpus.

What really matters is how frequent are the words in your corpus compared to a standard corpus? So for that, you have to compare these frequency lists against the frequency lists, for example, for the British National Corpus. And that's what terminology does, whereas you can, if you want to, just get out word frequency list from your corpus or multi-word expressions from your corpus.

You can also look at collocations, words which appear together quite commonly. You can look at a word sketch difference. So for a particular word, you can compare-- you can get the collocations. Or for two words like little and small, you might think they mean the same thing, but there are certain collocations with little which don't come with small and vice versa.

If you have a collected corpus over several years-- so for example, the Brown Corpus in America was collected for the 1960s and again for the 1990s and again for the 2010s. And you can see trends or changes in usage over time. So mouse, for example, in the 1960s means something a bit different from mouse in 2010.

And you can also, within Sketch Engine, do part-of-speech tagging. That is, work out, for each word in your corpus, mark it as being a noun or a verb or an adjective or whatever. You can also do lemmatization. So for example, dogs is made up of dog and the plural s. And you can do that not just for English, but for other languages too.

Finally, I want to look at web as corpus research. So Sketch Engine, you can use it to collect and analyze text data yourself. At Leeds, with Leeds University, we've done this for a very large corpus analysis. And there's also the ACL special interest group. We'll come back to that later.

Well, let's try it by actually going to the website. Let's have a look at Sketch Engine and see where this gets us. Now, I'm going to try it live today. This may work. It may not work tomorrow, but we'll see. If I click on that, then it opens up a web browser page. I've got my Chrome browser open on my MacBook. Here is the Sketch Engine web page. Lots of pretty pictures. There's videos, lots of videos.

Some of these are on YouTube, so you have to be logged in. I should have said, yes, you have to make sure that you connect via the VPN. So I have now connected via VPN. I can disconnect, and it won't work. So what I need to do is to connect, make sure that my VPN is connected. And here we are, connecting again. OK, once it's connected, then YouTube videos and Sketch Engine and stuff should work.

And it shows you what things Sketch Engine is used for. It says they've got lots of very large corpora for Arabic, Chinese, Japanese, and most languages you can think of. Here's some of the features, the ones I just looked at, and it covers a lot of different languages and scripts. So Arabic and Chinese don't use the same alphabet as English, but this system could handle that.

Here's some important people in linguistics and what they've said about how Sketch Engine is very good. There's a page showing prices where luckily you, as students, don't have to pay anything. But if, after this, you want to use Sketch Engine later on, there's a 30-day free trial license, or you may get your company to pay for you. Here's some companies that have used Sketch Engine for building dictionaries, for example.

OK, let's go back up to the top. We want to log in. If I click on Log in, that takes me to a page where I'm expected to log in. I don't have a use-- well, I don't want to log in this way. I want the institutional log in, and then I want to type in-- let's see. Leeds University. There's lots of universities in Leeds. We want University of Leeds. That's us. OK.

Now I'll then log me in via my university log in. Well, I've logged in already, so we won't go there. You will have to type in your username and password for Leeds University to log in. It's just warning me all accounts that my institution will close down on June 25, 2022. So the university has a license up until June 2022, so we're OK.

OK, here's various stuff. Here's a dashboard. I haven't selected a corpus yet. Let's go back to Select Corpus. Well, here's some corpora that I've done before or looked at recently. And let's create a new corpus. So our new corpus-- OK, I'm going to collect a corpus in the field of, let's say, data mining. Why not? And we're going to call the corpora Datamining. Datamining.

OK. It's a single language corpus, not a multilingual corpus. The language is English. We'll stick to that. Let's have a short description. This is a Trial corpus on data mining. It's just a test to see if it works. There's a whole load of extra features here, but I'm not going to bother with that just yet. So let's click on Next.

Now, I've got two options. I can either upload my own corpus if I have them. Well, I don't, so I'm now going to find text on the web. What do I want? You've got to give it a folder name. I'll just call it Web1 for now by default. There's a whole load of search settings. Well, what search settings are there? Let's have a look. Well, let's do the standard settings. So I won't do very much.

But one thing I do have to specify-- not seeing it. OK, I do want that. It says here, type at least three words or phrases. So I'm going to use web search. Web search means I'm going to give it some search terms to do with data mining. It will then use web search-- it will use a web search engine to find web pages which matches search terms, and it will download those texts.

So it's important I choose, it says, at least three words or phrases. Well, let's try data mining. That's a phrase. And return. And then what else? Let's say data and mining just to try it out. Those aren't very-- I mean, you probably want some more. And those aren't very clever, but whatever you are doing. Let's try those to start off with.

Now, we'll see the Go is-- before, it was blacked out because I couldn't start without doing at least that. There are other settings you can do too, but I'm not going to bother with those. When you do experiment with this, try some of the other settings that make changes. But let's just now press Go and see what happens. And what happens is it comes up with a whole list of web pages which the search engine matched my search terms. And not too surprisingly that these are to do with data mining.

First one is waikato.ac.nz.weka. So this is the Waikato University web page for the WEKA toolkit, which does data mining, so those are OK. I can, if I want to, deselect some of these because I don't think they're appropriate for some reason or other. But I'm not going to bother. I'm just going to say Go. Let's collect them all.

And it's now-- it's going to the web page. It's downloading the web pages and extracting from each web page the text. It's throwing away the images, tables, and what we call boilerplate. That is, on a web page, there's usually some sort heading at the beginning and some sort of-- at the bottom somewhere it will say how to contact the web owner, who owns the web page, links to other related web pages, stuff like that we don't really want.

What we want is the text which we hope will be about data mining. So it's taking some time to get rid of all this. And you'll find that if you have quite a few search terms or lots of web pages, this bit may take a while. But I've only done a simple example. So we've done that. You can say get to

know your corpus in various ways. Turn that off. Sorry about that. OK. Let's go to the corpus details and statistics.

Well, we've collected 29,000 tokens or 24,000 words. Was a subtle difference between tokens and words. We saw this in a previous lecture. So this is not a very big corpus, about 25,000 words. Not huge. But it will do to start off with. OK, we can go to Manage Corpus if you want to do some more stuff. We can make it bigger. We can do various other things with it.

And the other thing, going back to the dashboard, this is the dashboard for the data mining corpus we've got. Well, what we want it to do is get out some terminology, some key words from this corpus. So the key words extraction, well, again, you can just press the Go button and it will take a while.

But that's because what it does is it compares your corpus against some standard or reference corpus. And the default reference corpus is the English Web 2020, or enTenTen. That's 10 to the power of 10, or 10 billion words collected from the web. And that's fine because it's a very large representative corpus. But because it's very large, it will take quite a long time.

So let's look at the Brown Family instead. The Brown Family is one million words only from the 1960s and from the 1990s and again from 2010s. So it represents English over a time period, but it's a fairly small sample, so I've chosen that just to make the thing work faster for this example. OK, so this is words or phrases in my new data mining corpus which are significantly more frequent than in the Brown corpus that is in American and British English from '60s, '90s, and 2010s.

And we see that there are quite a few words like mining, funny words like doi. Well, what do these words mean? Well, you have to figure it out. Disproportionality. I'm not quite sure these are particularly data mining words. And you might want to look these up.

These are probably words which occur not very commonly at all, even in the data mining corpus, but they don't occur at all or maybe only one occurrence in the Brown corpus. So if you're taking your reference corpus, a small corpus, then you won't get a very good example. Statistically, not very significant meanings. This is the reason for choosing a large corpus as your reference corpus. But we've done this to make it work faster.

What about multi-word terms? Well, here, we're better. Data mining, data mine, adverse event, safety report, text mining. So more of the multi-word terms. So typically, in a domain, in a specialist domain, the multi-word terms are good indicators of that domain, whereas the single word terms, less so. You probably want to find multi-word terms for your domain.

OK, I think I've had a quick look at Sketch Engine and used it to collect corpus and extract some things from the corpus. So let's go back to our PowerPoint presentation. So we've had a quick look at Sketch Engine. Let's have a look at Leeds Internet Corpora. I won't spend that much time on this.

But here is the website, and you can query the corpora in much the same way as Sketch Engine queries the corpora. And you can get a concordance out. And again, most of these things, these are things that Sketch Engine could do for you. The real reason for looking at this web page is it has some information on open source development of large corpora. So if you want to collect a very large corpus of, let's say, a billion words or half a billion words from the web, then you have to collect not just two or three keywords.

You typically want to have about 500 keywords. And the keywords should be not for your particular domain, but representing the English language as a whole. So these are fairly frequent words, but they're not function words like the, of, and and, but they're typical words like picture, extent, raised, and so on. And we have these examples for-- it's the examples for English. Above, account, act, activities.

These are not particularly interesting words, but they're not very rare words either. And we've got the same thing for German, Russian, Chinese, and other examples. Having got this list of 500 words, then you use something like WebBootCaT to produce a list of about 5,000 to 6,000 queries. It takes combinations of these words, let's say four words from the list, and creates them as a query, sends them to the search engine, and back you get.

And here, we have some examples of the query sent out. There's topples of words. Then you download the URL produced by Google or whatever the search engine is. And here, we have the URLs, the web pages. You then have to-- what WebBootCaT then does is it does some post-processing, extracts the text, and gets rid of the rubbish you don't really want.

And then finally, you get some outputs, and you get some frequency lists from those corpora, which you can then check to make sure that they really are sensible. So what this page tells you is more about how you go about using something like WebBootCaT to collect a billion word corpus rather than just a small sample corpus that you're looking at.

OK. Let's go back to page here. What've we got? Next thing I want to look at is the Special Interest Group. Whoops. Can I click on this? We'll open the web page for the Association of the Computational Linguistics Special Interest Group on Web As Corpus. And the reason for doing this is just to point out this is a very good source of information because they have meetings or conferences pretty much every year or every other year sometimes.

And you can look at the proceedings, and the proceedings will contain research papers from all that were presented on Web As Corpus research. So in general, if you want to find out information about a topic area, yes, you can use Google, or you can use Google Scholar to find web pages, or you can look at the Association of Computational Linguistics Special Interest Group on your particular subject area, and you may find lots of conference proceedings in that topic area.

OK, that's enough for that. Oh, yeah. I did finally want to remind you that if you're going to use this, you have to use the VPN. And on that page, there was a link to the VPN. OK, so in this video, you've introduced Sketch Engine as a web tool for corpus linguistics. You've been introduced to WebBootCaT within Sketch engine as a way of collecting a corpus from the web. You can use it yourself to collect your own corpus and then to extract a list of terminology from that corpus.

Mainly the multi-word expressions are the important ones for your domain. You can look at Leeds internet corpora if you ever want to collect a billion word corpus, and this gives you more information about how to go about collecting seed words for this and then using those. Or if you want to do some research in a particular area, then look at the ACL Special Interest Group for that area. You should have read or be reading or about to read Adam Kilgarriff's paper on the search engine because it's got more examples in it. And look at the websites.

[END]