I’ve recently been rereading *Ulysses* by James Joyce and have been inspired with a computational question. *Ulysses* is a book that is often read alongside a lot of supplementary material, whether due to the novel’s difficulty, or perhaps simply due to the enormous amount of supplementary material available with claims to being the authoritative guide to the novel.

Joyce himself seemed to fall into the trap of feeling the need to provide a guide to his novel. On [two](https://en.wikipedia.org/wiki/Linati_schema_for_Ulysses) [separate](https://en.wikipedia.org/wiki/Gilbert_schema_for_Ulysses) occasions, he offered neat tables associating each of the eighteen episodes with various characters or places from Homer’ *Odyssey*, with different colors, with times of the day, and even with human organs.

Here is an image of the Gilbert Schema, sent from Joyce to his friend Stuart Gilbert:

The question is, then, how well does a computer do at creating a schema for *Ulysses*?

If we follow Joyce’s model for schematizing, there are some ideal characteristics we’d like to have. The first is understanding *Ulysses* at an episode resolution. Ideally, we would be able to reduce each episode to a single word that not only summarizes an episode in some way, but that also distinguishes that episode from the other seventeen. There are a number of computational tools available for detecting important keywords in a text, but for this problem, I settled on something called the hypergeometric test.

# Why hypergeometric?

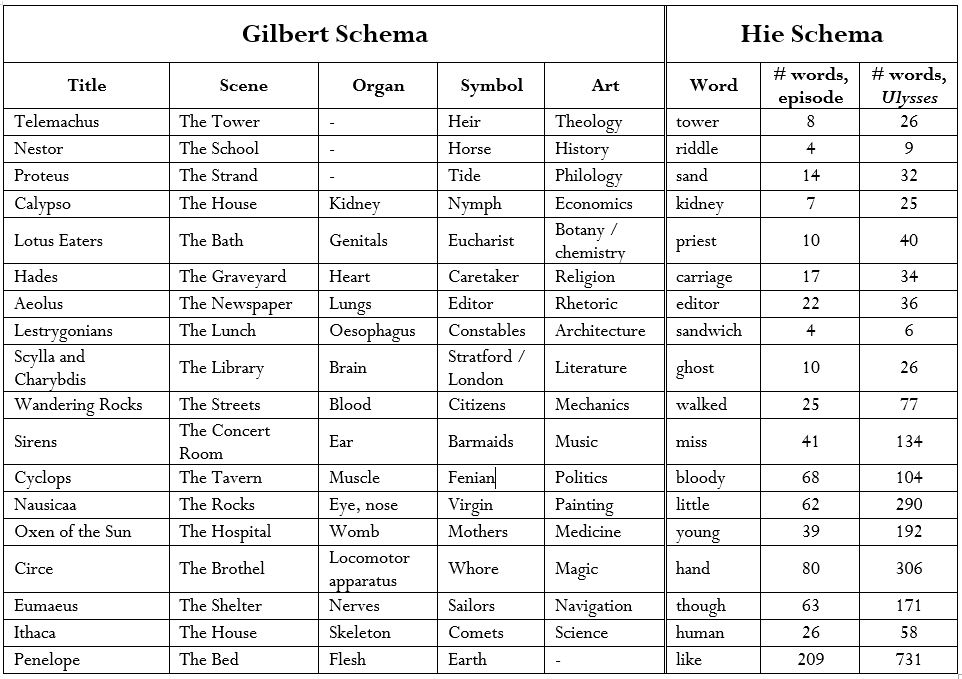
To visualize what a hypergeometric test does, consider a ball pit with a thousand balls, where only ten balls in the pit are painted red. Say your friend gives you a sample of twenty balls from the pit. You look in the sample and find nine red balls. You’d suspect, since there were so few red balls in the pit (only ten in thousand), that your friend has deliberately preferred the color red when sampling balls. The number of red balls in your friend’s sample seems a lot more than would be expected if you drew ten balls by chance. A hypergeometric test is a statistical method to confirm this intuition. Simply put, a hypergeometric test can be used to determine if a sub-population is overrepresented, or underrepresented, in a given sample of the entire population.

We can think of *Ulysses* as a bag of words, where each episode is a different (nonrandom) sample of those words. Words that are more important to a given episode will be overrepresented. Using a hypergeometric test, we can get an overrepresentation score for each word of each episode of *Ulysses*. A hypergeometric test works well since we are not only getting important words for a given episode, but also words that distinguish an episode from the rest of *Ulysses*.

Something important to consider is that these hypergeometric tests are not independent, since overrepresentation of a word in one episode will necessarily involve underrepresentation in other episodes. Still, we can take into account the many, non-independent hypergeometric tests we are doing to determine all the words in each episode that are significantly overrepresented. And to create our schema, we can simply pick the word that is the most overrepresented in each episode, according to the hypergeometric test score.

# Introducing the Hie Schema

So, how well did the computer do? It seems remarkably well, with a lot of interesting choices. Below are a few columns from the Gilbert Schema, with the final column coming from the computer’s schema (which I have taken the opportunity to name after myself):



All of the words in the Hie Schema are significantly overrepresented within their corresponding episodes. Many of the words picked out by the hypergeometric test are also in the Gilbert Schema: “tower,” “kidney,” and “editor.” There are also some that make a lot of sense, like the word “sand” being associated with the famous Proteus episode, where Stephen Daedalus walks along Sandymount Strand, “sandwich” being associated with the Lestrygonians episode, where Leopold Bloom has a gorgonzola cheese sandwich and a glass of burgundy for lunch, and “miss” representing the sirens episode.

# A Humanistic Catechism

The two most interesting words, though, are those of the last two episodes. The penultimate episode is the mock catechism, that was supposedly Joyce’s favorite. The most overrepresented word here is “human.” A quick search through uses of the word “human” in this episode reveal that the catechism’s questions and answers are highly preoccupied with corporal and carnal matters:

…the parts of the **human** anatomy most sensitive to cold being the nape…

…he was reluctant to shed **human** blood even when the end justified the means…

…consequent extermination of the **human** species, inevitable but impredictable…

…of the universe of **human** serum constellated with red and white bodies…

…the **human** organism, normally capable of sustaining an atmospheric pressure of 19 tons…

…fells of sewer rodents, **human** excrement possessing chemical properties…

…the 70 years of complete **human** life at least 2/7ths, viz., 20 years passed in sleep…

And in one particularly dense question:

He believed then that **human** life was infinitely perfectible, eliminating these conditions?  
There remained the generic conditions imposed by natural, as distinct from **human** law, as integral parts of the **human** whole: the necessity of destruction to procure alimentary sustenance: the painful character of the ultimate functions of separate existence, the agonies of birth and death: the monotonous menstruation of simian and (particularly) **human** females extending from the age of puberty to the menopause: inevitable accidents at sea, in mines and factories: certain very painful maladies and their resultant surgical operations, innate lunacy and congenital criminality, decimating epidemics: catastrophic cataclysms which make terror the basis of **human** mentality: seismic upheavals the epicentres of which are located in densely populated regions: the fact of vital growth, through convulsions of metamorphosis from infancy through maturity to decay.

In Joyce’s catechism, the world is mediated not through the language of the divine or the unseen, but rather through the physical, the bodily, and specifically the human. The catechism becomes comic but also deeply subversive by elevating a normal human, like Leopold Bloom, to the dogmatism and stylistic seriousness often only reserved for subjects of eternal significance.

# Molly Bloom and Valley Girl Stylistics

Molly Bloom’s monologue at the end of *Ulysses* is represented by a different word: “like.” Once again, looking at the uses of the word “like” reveals an important stylistic mechanism specific to Molly’s version of stream of consciousness, as opposed to Stephen’s or Leopold’s. (Try it yourself! Go to a digital version of *Ulysses* like [this one](http://www.online-literature.com/james_joyce/ulysses/18/) and CTRL-F for instances of “like.”)

In Stephen’s monologue, for example, the “stream” of consciousness is more like a chain of individual thoughts, each separated from the other by a comma or by a period. Molly’s monologue is different; it works more like an actual stream, consisting of phrases joined together by the word “like,” making the association from one thought to another more explicit and smoothing out the monologue into more or less a single syntactic unit.

These stylistic differences unique to Molly are representative of the way Joyce genders his characters’ thought. Stephen and Leopold think with a throbbing, rhythmic firmness; Molly thinks with a flowing, windy passivity.

# What’s the use of schemas?

Finally, a few words about schematizing *Ulysses*. I tend to agree with the criticisms that such schemas oversimplify the novel to a few words or set up spurious comparisons to Homer’s epic. However, the popularity of schemas, as well as heavily annotated versions of *Ulysses*,is something worth thinking about.

One way of understanding *Ulysses* is by looking at the trends in urbanization during the early twentieth century, during which cities became largely and immensely more complicated. For example, here is a map of the growth of London:

Urban sprawl happened more or less organically; block after block could be added, making the city seem almost infinitely expandable. The same is true for stream of consciousness, especially that of the early Leopold Bloom chapters. Joyce would add fragment after fragment, growing the size of these chapters and making them seem more and more complex. Eventually, navigating this kind of writing would be much like navigating urban sprawl, beginning on the broad boulevards of external, conscious stimuli before wandering through narrow alleyways of the internal and the subconscious, before emerging suddenly, once again, on a wide, open street.

Navigating an increasingly complex city required a street map, like the one of 1920s Dublin below:

Such maps are necessarily simplifications of the city. They include landmarks that help the city dweller orient him or herself. The same function is fulfilled in *Ulysses* by the schema or the annotated guidebook. *Ulysses* can be read as an attempt to capture the increasing complexity of the modern world in a novel. That is why a first-time reader of *Ulysses*, without the aid of a schema, might feel like a villager placed in the middle of New York City, without a map.