Manual markup exercise

As an exercise to introduce some concepts in content mining we will do a hands-on exercise in which we manually mine data from a scientific paper. The paper can be found in the GitHub repository for this workshop and paper copies will be available.

- You will be provided with six highlighters, and a paper copy of a scientific article. No knowledge of the specific subject matter is required for this exercise.
- Get into small groups (of up to six people) and each choose your highlighter colour(s) for marking **entities** - see below for the colour-code.
- Mark up the first 10 previously unseen entities on each page, starting at the top. Swap pages within your group so each is exposed to all of the markers.

Entities

An entity is a complete nounal word or phrase which can/could be found in a systematic taxonomy or definition. Thus “1984”, “Pleistocene”, “5 minutes” could be marked as dates/times, while “recently”, would be too fuzzy. Where there is ambiguity, try to consider whether others would vote the same way.

For each entity type, try to formulate rules that you used and how they could be computerised. Examples are:

- syntactic context: “the *foo* lives in *bar”*
- lexical makeup of words
- style or font
- lookup in authorities (*e.g.* Wikipedia)
- computability - can the entity be analysed algorithmically?

Please add your answers and thoughts to the etherpad

Types of entity

- YELLOW: **species and genera** - use only precise terms (*e.g.* omit generic terms such as “pets”)
- ORANGE: **places** - anything that resolves to a usefully defined point or region (*e.g.* “London”, latitude+longitude, not “further north”). Exclude metadata.
- PURPLE: **dates/times** - resolvable to an ISO 8601 date time or an ontologically supported concept (*e.g.* “Eocene”, “equinoctual”, not “recently”). Exclude metadata.
- PINK: **identifiers** - formal codes given by authorities (*e.g.* PMIDs, Genbank IDs).
- CYAN: **bibliographic and academic metadata** - including rights, grants *etc*. Indicate precise lengths of phrase and constant/variable components.
- GREEN: **chemical compounds** - resolvable in PubChem, or of generic chemical form (*e.g.* not “diesel oil” or “butter”).

Precision and recall

When your group has finished, swap sheets with a neighbouring group. Critique their markup, drawing a circle around places that you disagree - either for an omission (false negative), or a wrongly marked entity (false positive).

*Note: there is **NO** absolutely right or wrong answer, there is only inter-annotator dis/agreement. The better the agreement, the higher scores we should expect from programmatic output.*

This process is tedious and generally impossible without open material. Humans must:

- write and agree the rules - we wrote 31 pages of rules to interpret chemistry
- have several people independently mark up a test corpus and then calculate the inter-annotator agreement: we got 92% for chemistry
- write a program and train it against a training corpus (more work!)
- use a validation corpus (even more work!) to test how well the program performs