**Taxonomy & other Puzzles**  Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This is NOT a test! There is no clear or correct answer for many of the items below. The idea is to show you some of the tasks and data that we deal with and to provoke some thought and some discussion of how we might analyze them.   
**Note:** Your answers can be longer than the tiny spaces provided. Be prepared to explain why you gave each answer and what patterns you saw in each problem.

1. **Professional skills**

People list unusual terms under “skills” as part of their professional profile on LinkedIn; sometimes these skills are either not relevant or might be criminal activities. How would you classify the “skills” below? How did you decide?

|  |  |  |
| --- | --- | --- |
| Term | A professional  skill? (yes/no) | Why? / Why not? |
| Headaches |  |  |
| Beer |  |  |
| Lingerie |  |  |
| Fraud |  |  |
| Teens |  |  |
| Chef |  |  |
| Microscope |  |  |
| Workplace violince |  |  |
| Wigs |  |  |
| Child labor |  |  |

1. **Job titles**

People put a wide variety of things in the job title fields on their LinkedIn profile. Some are clearly job titles like “Marketing Manager”; others are not so clear. How would you classify the examples below? How did you decide?

|  |  |  |
| --- | --- | --- |
| Term | A clear job title? (yes/no) | Why / Why not? |
| Captain |  |  |
| Architect |  |  |
| Sales |  |  |
| Freelancer |  |  |
| Arborist |  |  |
| Resident |  |  |
| Crew member |  |  |
| Profesional |  |  |
| Associate |  |  |
| Six Sigma Black Belt |  |  |

1. **More job titles**

What are the most important *kinds* of information in job titles like these? I.e., what are the obligatory and optional parts of a good job title? Don’t analyze these titles one by one; **give a single overall response**.

|  |  |
| --- | --- |
| Term | Answer |
| adjunct professor (part-time), department of biology |  |
| retired high school english teacher |  |
| freelance blogger |  |
| master plumber |  |
| senior marketer |  |
| assistant to genral director |  |
| project manager, pmp |  |
| professional actor |  |
| warehouse worker |  |
| part-time library specialist |  |

1. **Classification**

A part of our work is to classify terms that appear on member profiles. For example, we’d like to find people based on which **tools** they know how to use. How would you classify the terms below?

|  |  |  |
| --- | --- | --- |
| Term | Is this a tool? (yes/no) | Why? / Why not? |
| Hammer |  |  |
| Scalpel |  |  |
| Ballet shoes |  |  |
| Crossword puzzle |  |  |
| Chair |  |  |
| Algebra |  |  |
| HTML |  |  |
| Nail |  |  |
| Drill bit |  |  |
| ISO 9001 |  |  |

|  |
| --- |
| How would you define the concept of a *tool* to summarize your responses? |
|  |

**5.** **What’s important?**

What are the most important pieces of information we can use to match a job description like this one to a LinkedIn member? What *kinds* of information are they? **Make a list of kinds of information below the job description.**

|  |
| --- |
| **Taxonomy Analyst** We are seeking extraordinary talent to work alongside machine learning researchers, software engineers, data experts, and product managers to build out LinkedIn’s Professional Knowledge Graph -- our knowledge bases about economic entities like professions, companies, and skills. This work drives all of the products at LinkedIn so that we can create economic opportunity for professionals everywhere.  Our ideal candidate will be passionate about applying his or her expertise to increase the accuracy and coverage of the machine learning systems that we use to classify information from jobs and members. This position is a full-time contract role at our office in Sunnyvale, CA. Responsibilities of this role In this role, you will create, enrich, and validate our knowledge bases and taxonomies by creating new knowledge and evaluating how that knowledge helps in practice. These include tasks like the following: Knowledge Creation  * Identifying systematic relationships between concepts such as job titles, skills, certificates, etc. * Researching, classifying, and annotating the phrases, terms, and data in messy, user-generated data * Research new entity categories and best practices for organizing them * Leverage external resources like Wikipedia and standards to enrich and organize our data  Evaluation  * Create instructions for outsourced evaluations, for example with CrowdFlower * Evaluate the output of machine learning classifiers and internal tools  Basic Qualifications:  * Strong problem solving and conceptual thinking skills * Analytical with high attention to detail * Excellent written and spoken communication skills in English * Willingness to learn technical concepts and new tools and a keen interest in technology * Comfortable working with unclear classification rules or on unfamiliar topics * Ability to prioritize and explain decisions * BS/BA Degree in Linguistics, Computational Linguistics, Library and Information Science, Philosophy, Experimental Psychology, Mathematics, or related discipline. * 1+ years of experience with taxonomy work, search, controlled vocabularies, and/or linguistic annotation  Preferred Qualifications: ● Ability to understand and apply complex task-specific instructions with minimal supervision   * Understand the tradeoffs involved in making classification decisions   ● Practical experience with the messiness of user-generated content  ● Knowledge of semantics, dictionary building, controlled vocabularies, taxonomies, and/or ontologies  ● Knowledge of tools for natural language processing and/or CrowdFlower  ● MS or PhD in Linguistics, Computational Linguistics, Library and Information Science, Philosophy, Experimental Psychology, Mathematics, or related discipline. |

**6. Questions about questions**

**Target Question:** *How likely is this talent pool to relocate to a new geography?*

|  |  |
| --- | --- |
|  | Does each of these questions ask for the **same exact information** as the Target Question above? If not, explain why not. |
| Would these people move to a new city? |  |
| What's the chance this talent is willing to move to a new location? |  |
| How many of them will relocate? |  |
| Can he relocate? |  |
| What proportion of them would change locations? |  |

**The information I want:** *The number of job postings available for people in this group*

|  |  |
| --- | --- |
|  | Does each of these questions ask for the **same exact information** as the information I want above? If not, explain why not. |
| How many jobs per candidate for this talent? |  |
| How many job openings are there for each professional in this talent pool? |  |
| How hot is the job market for this talent? |  |
| What is the demand for this kind of talent in California? |  |
| What is the percentage of jobs for this talent pool? |  |

**7. Recap**

What are the hardest things that you noticed in all of these puzzles, from 1 to 6?

|  |
| --- |
|  |

Extra, optional challenge: **Similar Terms**

Sometimes terms look very similar and we have to decide whether to merge them as if they were synonyms or to split them as separate concepts. For that, we use other data such as skills that appear frequently with a given term (*top skills*) or job titles that appear frequently with that term (*top titles*).

**For each pair of titles below, say whether we should merge the pair or split the pair. Then explain why.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term Pairs | Top skills | Top titles | Merge  or split? | Why? / Why not? |
| Biotechnology Industry | microsoft office, biotechnology, pharmaceutical industry, life sciences, project management, management, molecular biology | consultant, director, president, founder, chief executive officer |  |  |
| Biotechnology | research, cell culture, medical devices, life sciences, pharmaceutical industry, molecular biology, biochemistry | president, chief executive officer, director, owner, consultant |  |  |
|  |  |  |  |  |
| Computer Gaming | microsoft office, computer hardware, microsoft excel, photoshop, online gaming, microsoft word, customer service | founder, chief executive officer, owner, student, cofounder |  |  |
| Computer Games | game development, gameplay, mobile games, xbox 360, casual games, game design, video games | game designer, founder, owner, cofounder, chief executive officer |  |  |

Extra, optional challenge: **Programming Task**

Write a script that solves the following two problems and send it by email along with your output files and your other responses. Be prepared to explain what challenges you faced and why you solved them the way you did.

1. Assume you have the following text in a .txt file:

**###CLINICAL NURSE I**

**\*\*\*\*Research Nurse -- Neurology**

**\*\*\*Sign on Bonus\*\*\*Staff Nurse - SJHC Telemetry -Nights F/T**

**Clinical Nurse-E2 ICU Med/Surg/Neuro/Trauma ICU 0.9 rotating (34255) Â»**

**Care Giver \\ Home Health Aide \\ CNA**

Process the text such that it matches the output text below.

Print the output to a different txt file with the original string in one column and the processed string in the next column on the same line.

**clinical nurse i**

**research nurse neurology**

**sign on bonus staff nurse sjhc telemetry nights f t**

**clinical nurse e2 icu med surg neuro trauma icu 0.9 rotating 34255**

**care giver home health aide cna**

1. Get unigram frequencies of each word in the processed text.  
   Output each unigram along with its frequency (in separate columns) to a new csv file using the csv library in Python (<https://docs.python.org/2/library/csv.html>)