1. (14 points) Assume that simple term frequency weights are used (no IDF factor), and the only stopwords are: “is”, “am” and “are”. Compute the cosine similarity of the following two simple documents:

(a) “precision is very very high”
(b) “high precision is very very very important”
2. (14 points) Assume that an IR system returns a ranked list of 10 total documents for a given query. Assume that according to a gold-standard labelling there are 5 relevant documents for this query, and that the only relevant documents in the ranked list are in the 2nd, 3rd, 4th, and 8th positions in the ranked results. Calculate and clearly show the interpolated precision value for each of the following standard recall levels: \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\} for this individual query.
3. (14 points) The table below shows the final ranked list of results for an IR search together with their continuous human-rated relevance values. Assume the table contains all documents with non-zero relevance. Compute the values of the DCG and NDCG evaluation metrics for each value of \( n \) and add them to the table. Complete the second table to show the idealized DCG (IDCG) values.

<table>
<thead>
<tr>
<th>( n )</th>
<th>doc</th>
<th>relevance (gain)</th>
<th>DCG</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D23</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>D78</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>D90</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>D17</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>D78</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( n )</th>
<th>doc</th>
<th>relevance (gain)</th>
<th>IDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. (13 points) Write a Perl regular expression (regex) for matching the final line in a US Postal address in Texas or California. Assume that it consists of a city name of one or two alphanumeric words followed by a comma and then any amount of optional whitespace, followed by one of the two-letter state abbreviations (TX or CA) followed by some whitespace and then a 5 digit zip-code with an optional “plus four” digits introduced by a hyphen.
5. (12 points) Assuming Zipf’s law with a corpus independent constant $A = 0.1$, what is the fewest number of most common words that together account for more than 18% of word occurrences (i.e. the minimum value of $m$ such that at least 18% of word occurrences are one of the $m$ most common words).
6. (12 points) Draw the DOM tree for the following XML document:

```xml
<db>
  <customer>
    <name>
      <firstname>John</firstname> <lastname>Doe</lastname>
    </name>
    <phone>
      <areacode>512</areacode> <number>471-9558</number>
    </phone>
    <purchases>
      <item>
        <camera>
          <type>Canon digital</type> <price>200</price>
        </camera>
      </item>
    </purchases>
  </customer>
</db>
```
7. (21 points) Provide short answers (1-3 sentences) for each of the following questions:

What is the difference between database management and information retrieval?

Why is Euclidian distance not a good metric for judging the (dis)similarity of documents in vector-space retrieval?

What is functional role of the IDF factor in standard term weighting?

How does stemming typically affect recall? Why?

Define “pseudo relevance feedback”.

Why does thesaurus-based query expansion typically not work very well?

On what type of plot does a power law result in a straight line? What is the slope of the line (in terms of the parameters of the power law)?

(Extra credit) A “portmanteau” is word that combines parts of two or more words into a single word such as “bit” for “binary digit” and “regex” for “regular expression.” Who coined the term “portmanteau” (in this sense) and who coined the term “bit” (as a measure of information)?

(Extra credit) Who is the only person to ever be awarded both a Nobel Prize and a Turing Award?