Programming Assignment: Article Similarity Calculation

This assignment focuses on fundamental text processing and similarity calculation techniques using Python. You will implement a solution to analyze a given CSV file containing articles, clean their content, build a global bag-of-words representation, and then calculate the cosine similarity between these articles. The entire process should be implemented using only Python's built-in `csv` module and the `numpy` library for numerical operations.

Assignment Description

Your task is to write a Python script that performs the following steps:

1. **Read the CSV File**: The input will be a CSV file named `articles.csv` with three columns: `id`, `title`, and `content`. You must use Python's `csv` module to read this file. No external libraries like `pandas` are allowed for this step.

2. **Clean Article Content**: For each article's `content`, perform the following cleaning operations:

   - Convert all text to lowercase.
   - Remove punctuation (e.g., commas, periods, apostrophes, exclamation marks, question marks, etc.).
   - Remove numerical digits.
   - Tokenize the cleaned content into individual words.

3. **Build Global Bag-of-Words (BoW) Vocabulary**: Create a **single, global vocabulary** of all unique words found across all articles. This vocabulary will serve as the basis for representing each article.

4. **Build Vector Representation**: For each article, construct a vector representation based on the global BoW vocabulary. Each element in the vector will correspond to a word in the global vocabulary, and its value will represent whether or not that word appeared in the specific article (0 or 1).

5. **Calculate Cosine Similarity**: Using the article vectors, calculate the cosine similarity between every pair of articles. You **must** use the `numpy` library for all numerical computations involved in calculating cosine similarity. Do not implement the cosine similarity formula manually without `numpy`.

   The formula for cosine similarity between two vectors A and B is:

$$\text{cosine similarity} = \frac{A \cdot B}{||A|| \cdot ||B||}$$

6. **Output Similarity Matrix to PKL**: Save the entire similarity matrix (a square matrix where `matrix[i][j]` is the similarity between article `i` and article `j`)

into a Python pickle file named `similarities.pkl`. This is the primary output of the assignment.

7. **Find Most Similar Articles**: Build a function that takes an `article_id` as input and returns the titles of the 3 articles with the highest cosine similarity to the input article (excluding the article itself), sorted by highest similarity to lowest.

Example Input (`articles.csv`)

```
id,title,content
1,The Rise of AI,Artificial intelligence is transforming industries
globally. Machine learning and deep learning are key components…
2,Future of Robotics,Robotics is advancing rapidly, integrating with AI
for autonomous systems. Human-robot interaction is a growing field…
3,Data Engineering, Data Engineering leverages tools and computation to
move massive amounts of data. Big data analytics is crucial…
```

Hints

- For cleaning, regular expressions (`re` module) can be very helpful, but not necessary.
- Use the ‘In’ key word in python for simpler code.

Good luck!