Define a Word as `{ALPHA}({ALPHA}-')*`

What's a "word?"

Let's start with this one.

Let's use `{ALPHA}({ALPHA}-')*`,
which means

- **start with a letter** (the first `{ALPHA}`)
- **followed by 0 or more** (the `*`)
- **letters, hyphens, or apostrophes**.

We Will Ignore Case

Should case matter?

Consider some examples…

"The" and "the" …

… are the same word? YES!

"Jack" and "jack" …

… are the same word? NO!

We're not going to solve this problem today.

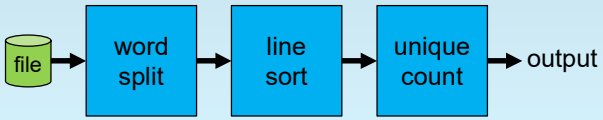
Let's choose to ignore case.

What Limits the Number of Words?

How many words are allowed?

Let's look at the pieces again…

- Word split only looks at one word at a time.
- The same is true of unique count.
- Line sort has to look at all lines.



We Limit Line Sort to 5,000 Lines

How many words are allowed?

Use dynamic allocation to allow as many words as fit in memory?

But we only know insertion sort ($O(N^2)$ time).

We'll limit the code to 5,000 lines/words.