



CLAIRVOYANT

Improved Capability at Lower Cost With Machine Learning, Cloud, and Open Source

October 17, 2018

Teaching students to write is a key challenge in higher education. Employers say they want universities to focus more on written communication skills, but universities are under increasing pressure to control costs. The challenge presented to Clairvoyant was to build a seamlessly-integrated writing feedback tool so students get rapid, high-quality feedback without driving per-student costs through the roof. This paper describes the challenges Clairvoyant faced, and the solutions we engineered to help our client succeed.

It's 11:30 pm and Lauren's breath gets a little quicker every time she reads a paragraph on her glowing laptop screen. Her hands get shakier each time she looks at her laptop's tiny digital clock. She volunteered to submit the final paper on behalf of her team, but with 30 minutes left until the deadline, she's wishing she had stayed quiet and let someone else do it. Everyone sent their work, and it all seems to be in order, but those last couple of paragraphs have some grammar problems and some of this just doesn't *sound* like the teammate she knows. If he stole that from somewhere, it will be the whole team's problem.

It's not unusual for students to be in Lauren's position. Universities have codes of conduct and policies governing plagiarism. They have systems in place that seek to detect and punish wrongdoers and ensure that everyone is playing by the rules. The problem is that enforcement-based approaches are all reactive, and they all work from the assumption that the only reason students include unoriginal content is to cheat. Students might still be learning the right way to cite their sources and format their references.

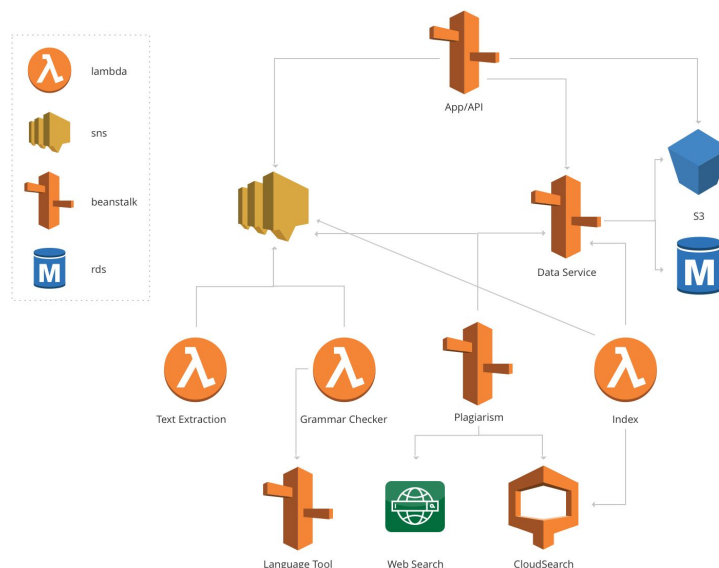
The ideal approach is to intervene early and provide support for students to do the right thing. Giving them a self-service writing assistant will help improve their understanding of properly-cited sources in the same place they receive feedback on their spelling, grammar and writing style.

That's the idea our client brought to Clairvoyant: let's build a general-purpose writing feedback tool to give students help with all aspects of their writing, including plagiarism checking. The ideal system would:

- Present all student writing feedback in one place
- Work in students' web browser with no software to install
- Integrate seamlessly with their Learning Management platforms
- Have a configurable grammar, style, and spelling checker
- Access a large database of content to check for unoriginal work
- Work faster than existing solutions on the market
- Have a low per-student cost

Clairvoyant blended custom application development with services from multiple cloud providers to choose a best-of-breed solution to each challenge.

STRUCTURE OF THE SOLUTION



FRONT END

Users send their papers to the system through their Learning Management platform's built-in assignment management features. They can check a document prior to sharing the work with their instructor. When they turn in their final draft, it will be run through the checker one more time for their instructor to see and add to the feedback. The whole feedback process takes less than two minutes and students can view the results by clicking a link right next to their assignment in the online classroom.

FEEDBACK GENERATION AND AGGREGATION

Spelling, grammar, and style feedback come from LanguageTool running a rule set customized for the US English-speaking university population we're serving. It's a good fit for this client because its rule set is very good out of the box and it gives them the tools they need to make it incrementally better over time.

PLAGIARISM CHECKING

The system handles plagiarism checking in multiple phases:

- Search for matches in an index containing millions of papers submitted to the university,
- Search for matches on the internet using Web Search APIs, and
- Analyze each result from both searches to see if it's really a match.

Faculty also have the opportunity to add their own comments to the feedback generated by the system, so students truly have a single report containing feedback on every aspect of their writing.

DELIVERY TO USERS

Once all the feedback has been aggregated, a custom report is generated to be shared with the student. Feedback is displayed as discrete items pointing to the text that they reference. In cases of suspected plagiarism, the original document is displayed side-by-side with the matching content and the relevant text is highlighted on each side.

TECHNICAL CHALLENGES

INTENSELY BURSTY TRAFFIC CREATES UNFAIR RETURN TIMES

For scalability, we used Amazon Web Services' Elastic Beanstalk and Lambda services. These components automatically scale all layers of the application as needed to ensure students have a high-performance experience.

Student traffic for this system is intensely clustered around assignment turn-in deadlines. Each component scales independently and automatically, but even the most rapid expansions still take time. When your traffic curve looks less like a ramp and more like a wall, you can get a lot of papers assigned to a small cluster in the time it takes to expand. This results in longer wait times for users who submit early in the ramp-up compared to the shorter wait times for those who submit just as new servers wake up.

The solutions were to pre-scale in preparation for known busy times and to switch to a queued approach. Rather than pushing traffic at a cluster and trusting it to scale, we have a bigger cluster just in time to pull documents in order. This prevents the original servers from over-committing, and results in documents being returned in close to the order they were received.

SEARCH IS EXPENSIVE

The costliest component of the system is the web search portion of the plagiarism check. In small numbers, web searches aren't expensive, but running a lot of them per document for a large population of students adds up quickly. We took a multi-faceted approach to limiting our consumption of web search while maintaining quality.

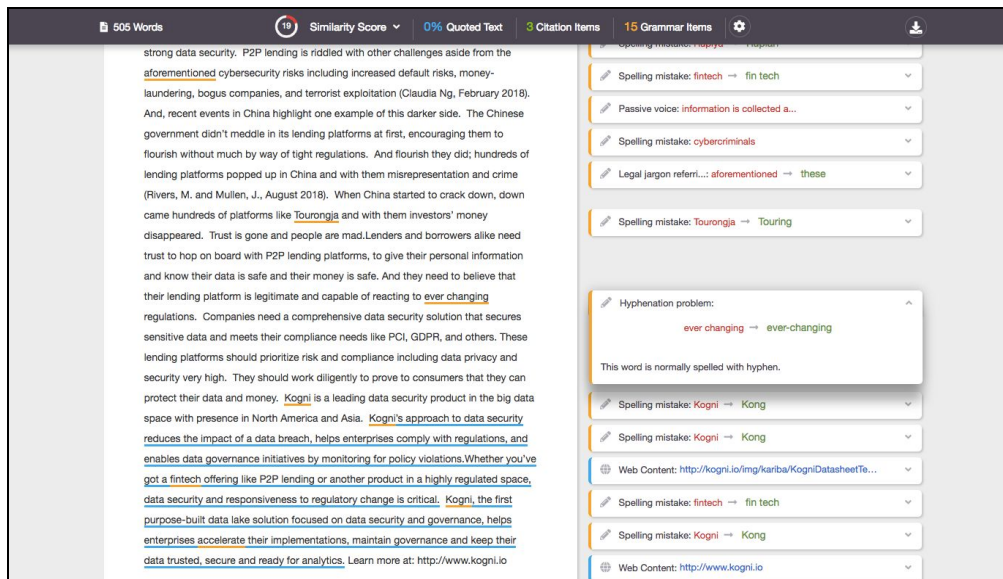
A lot of approaches are simple and don't come with a quality cost. About 1 in 20 documents submitted to the system are exact duplicates of documents that were submitted recently by the same author. It's inexpensive and safe to deliver the same report in response. We can also ignore text found in the assignment itself or in templates provided by instructors. That actually comes with a quality improvement as it avoids adding meaningless feedback to the report. We were able to effect a 25% reduction in search volume using lossless techniques like those.

A 25% reduction wasn't enough for our client. Out-of-control college costs have been grabbing headlines for years, and universities need tools with low per-student costs.

Working in our favor, we have a huge set of assignments submitted over a period of years. In addition, we have documents containing actual incidents of plagiarism discovered by university staff. We used machine learning to build a classifier capable of predicting the likelihood of plagiarism within a given segment of text. This approach allows the system's administrators to dial in the number of searches to attain the desired level of text coverage. Furthermore, they don't have to use a one-size-fits-all approach: introductory writing courses can be treated differently from upper division math courses. Using these techniques, administrators can achieve 5% to 50% reductions in search volume in exchange for reductions in duplicate text findings ranging from a few hundredths of a percent to a few percent.

MEASURING SUCCESS

When the client approached us, we worked with them to zero in on their definition of success. We need to know not only that we delivered software and people are using it, but that it's changing something meaningful for our business stakeholders.



The screenshot displays a plagiarism checker interface. The top navigation bar includes: 505 Words, Similarity Score (19%), 0% Quoted Text, 3 Citation Items, 15 Grammar Items, and a settings icon. The main document area contains text with several underlined segments. The right sidebar lists 15 grammar and spelling items, including: Spelling mistake: fintech → fin tech; Passive voice: information is collected a...; Spelling mistake: cybercriminals; Legal jargon referri...: aforementioned → these; Spelling mistake: Tourongia → Touring; Hyphenation problem: ever changing → ever-changing; Spelling mistake: Kogni → Kong; Spelling mistake: Kogni → Kong; Web Content: http://kogni.io/img/kariba/KogniDatashetf...; Spelling mistake: fintech → fin tech; Spelling mistake: Kogni → Kong; and Web Content: http://www.kogni.io.

THE IDEAL SYSTEM

By combining an extensible spelling, grammar and style product with a custom plagiarism checker and faculty feedback tool, we met the client's primary goals of presenting all writing feedback in one place, in the browser, direct from the Learning Management platform.

We started with a database of millions of papers, then used a machine learning approach to make intelligent use of our most expensive resources. This keeps our client's costs low and equips them with the tools to tailor system cost and performance for different populations.

- ✓ Present all student writing feedback in one place
- ✓ Work in students' web browser with no software to install
- ✓ Integrate seamlessly with their Learning Management platforms
- ✓ Have a configurable grammar, style, and spelling checker
- ✓ Access a large database of content to check for unoriginal work
- ✓ Work faster than existing solutions on the market
- ✓ Have a low per-student cost

The students also needed something fast. The university used to advise students that they should expect the old plagiarism checker to take an hour or more to return a report. With this new system, response times are less than two minutes, even during the busiest times. Students love that, particularly as they get close to assignment turn-in deadlines.

Now let's check in on Lauren. She has a lot of material for a group assignment, but one classmate has included some content that definitely needs grammar help, and some of it might be plagiarized. Lucky for her, there's help available right in the online classroom. She submits the whole document, but leaves it as a draft so it's not shared with her instructor yet. Less than two minutes later, she's looking at a feedback report. It has good suggestions for the grammar issues, and it turns out all the sources of possible plagiarism were already included in the References page. Her classmate didn't properly quote and cite a couple of things. There's plenty of time to clean that up and submit before midnight.

ABOUT CLAIRVOYANT

Clairvoyant is a global technology consulting and services company. We help organizations build innovative products and solutions using big data, analytics, and the cloud. We provide the best-in-class solutions and services that leverage big data and continually exceed client expectations. Our deep vertical knowledge combined with expertise on multiple, enterprise-grade big data platforms helps support purpose-built solutions to meet our client's business needs.

CLAIRVOYANT